



PROCEEDINGS
OF THE
ROYAL SOCIETY OF LONDON

SERIES A. MATHEMATICAL AND PHYSICAL SCIENCES

VOL 204



LONDON

Published for the Royal Society by the
Cambridge University Press
Bentley House, N.W.1

9 January 1951

*Printed in Great Britain at the University Press, Cambridge
(Brooke Crutchley, University Printer)*

*and published by the Cambridge University Press
Cambridge, and Bentley House, London*

Agents for Canada, India, and Pakistan: Macmillan

CONTENTS

SERIES A VOLUME 204

No. 1076. 22 November 1950

	PAGE
A discussion on detonation. By W. G. Penney, F.R.S. and others. (Plates 1 to 3)	1
A study of sensitized explosions. X. The kinetics of decomposition of chloropicrin and of the hydrogen-oxygen and hydrogen-chlorine reactions sensitized by chloropicrin. By P. G. Ashmore and R. G. W. Norrish, F.R.S.	34
Compressibility and absorption frequency of ionic crystals. By B. Szigeti	51
Kinetic studies on carbohydrates in alkaline conditions. I. The kinetics of the autoxidation of glucose. By C. H. Bamford and J. R. Collins	62
Kinetic studies on carbohydrates in alkaline conditions. II. The kinetics of the rearrangements of glucose and fructose in alkaline solution. By C. H. Bamford, Daphne Bamford and J. R. Collins	85
The thermal conductivity of tin, mercury, indium and tantalum at liquid helium temperatures. By J. K. Hulm	98
On the quantum theory of the elementary particles. I. Introduction and classical field dynamics. By K. V. Roberts	123

No. 1077. 7 December 1950

Wave equations in momentum space. By Maurice Lévy	145
An interferometer microscope. By J. Dyson. (Plates 4 to 7)	170
Paramagnetic resonance in chromic sulphate alums at room temperature. By D. M. S. Bagguley and J. H. E. Griffiths	188
Paramagnetic resonance spectra of five chromic sulphate alums at low temperatures. By B. Bleaney, F.R.S.	203
Thermal properties of potassium chromic alum between 0.05 and 1° K. By B. Bleaney, F.R.S.	216
The scattering of gamma-rays in extended media. I. Perpendicular incidence on a plane slab. By L. Cave, J. Corner and R. H. A. Liston	223
The velocity of propagation of electromagnetic waves derived from the resonant frequencies of a cylindrical cavity resonator. By L. Essen. (Plate 8)	260
A critical examination of the gravitational theory of the origin of cosmic rays. By P. J. D. Gething	278
Oxidation-reduction processes in cultures of bacteria. I. The reducing power of <i>Bact. lactis aerogenes</i> . By A. C. Baskett and Sir Cyril Hinshelwood, F.R.S. (Abstract)	295

No. 1078. 22 December 1950

	PAGE
On the stability of a fluid heated from below. By O. G. Sutton, F.R.S.	297
An experimental and thermodynamic investigation of the hydrogen-titanium system. By A. D. McQuillan	309
The scattering of gamma-rays in extended media. II. Back-scattering of gamma-rays from a thick slab. By J. Corner and R. H. A. Liston	323
The scattering of gamma-rays in extended media. III. Problems with spherical symmetry. By J. Corner, F. A. G. Day and R. E. Weir	329
Interferometric studies in diffusion. I. Determination of concentration distributions. By C. Robinson. (Plates 9 and 10)	339
Surface energy and structure of crystal boundaries in metals. By K. T. Aust and B. Chalmers. (Plate 11)	359
The viscosity of liquid helium between 2 and 5° K. By R. Bowers and K. Mendelssohn	366
The calculation of deflexions of the vertical from gravity anomalies. By A. H. Cook	374
Fermi-Dirac functions of integral order. By P. Rhodes	396
Theory of light absorption and non-radiative transitions in <i>F</i> -centres. By Kun Huang and Avril Rhys	406
Ultrasonic dispersion in organic vapours. By J. D. Lambert and J. S. Rowlinson	424

No. 1079. 9 January 1951

The invariant theory of isotropic turbulence in magneto-hydrodynamics. By S. Chandrasekhar, F.R.S.	435
The effect of an electric field on the viscosity of liquids. II. By E. N. da C. Andrade, F.R.S. and D. Dodd	449
Kinetics of the hydrogen/oxygen reaction. I. The explosion region in boric acid-coated vessels. By Sir Alfred Egerton, F.R.S. and D. R. Warren	465
Binding energy of the triton. By Tsi-Ming Hu and Kung-Ngou Hsu	476
Experiments on the reaction $T(d, n)^4\text{He}$. I. The 90° cross-section. By D. L. Allan and M. J. Poole	488
Experiments on the reaction $T(d, n)^4\text{He}$. II. The angular distribution. By D. L. Allan and M. J. Poole	500
The theory of the $T + D$ reaction. By B. H. Flowers	503
Orientation of fatty acid and soap films on metal surfaces. By J. W. Menter and D. Tabor. (Plates 12 and 13)	514
Structure of thin films of aliphatic esters and alcohols on metals. By J. V. Sanders and D. Tabor. (Plates 14 and 15)	525
Light distribution near focus in an error-free diffraction image. By E. Wolf	533
Interferometric studies in diffusion. II. Influence of concentration and orientation on diffusion in cellulose acetate. By J. Crank and C. Robinson. (Plates 16 and 17)	549
Statistical kinetics of Hevea rubber cyclization. By M. Gordon	569
Index	586

A discussion on detonation

CONTENTS

	PAGE
W. G. PENNEY, F.R.S. (THE LEADER OF THE DISCUSSION). INTRODUCTION	1
SIR GEOFFREY TAYLOR, F.R.S. SIMILARITY SOLUTIONS TO PROBLEMS INVOLVING GAS FLOW AND SHOCK WAVES.	8
H. JONES. THEORETICAL CONSIDERATIONS RELATING TO THE DETONATION OF SOLID EXPLOSIVES	9
W. M. EVANS. SOME CHARACTERISTICS OF DETONATION	12
R. M. DAVIES, J. D. OWEN, D. H. EDWARDS & D. E. THOMAS. PRESSURE MEASUREMENTS IN DETONATING GASES	17
C. A. ADAMS. TWO NEW STREAK CAMERAS	19
F. P. BOWDEN, F.R.S. THE INITIATION OF AN EXPLOSION AND ITS GROWTH TO DETONATION	20
A. R. UBBELOHDE. THE PHYSICAL CHEMISTRY OF DETONATIVE REACTIONS	25
J. S. COURTNEY-PRATT. A NEW PHOTOGRAPHIC METHOD FOR STUDYING FAST TRANSIENT PHENOMENA	27
A. R. UBBELOHDE. SOME COMMENTS ON THE CONTRIBUTION BY DR COURTNEY-PRATT	29
S. PATERSON. THE REDUCTION IN REACTION ENERGY IN CASES OF INCOMPLETE CONFINEMENT	30
J. TAYLOR. HIGH AND LOW DETONATION VELOCITY REGIMES IN CONDENSED EXPLOSIVES	30
O. A. GURTON. THE FADING OF DETONATION IN SOLID EXPLOSIVES	31
REFERENCES	32

(Discussion held 10 November 1949.—Complete MS. received 10 May 1950)

[Plates 1 to 3]

INTRODUCTION TO THE DISCUSSION

BY W. G. PENNEY, F.R.S.

The fundamental physical, chemical and mechanical processes which occur when a detonation wave passes through an explosive were imperfectly understood at the beginning of the recent war. As part of the scientific war effort in the British Commonwealth and in the United States of America, many theoretical and experimental studies were made of detonation processes. Much of the work has recently been declassified and some has been published. Several centres of research in this country and elsewhere are vigorously continuing with these studies. As later papers will show, the quality and general scientific interest of much of this work was considered sufficient to form the basis of a Discussion of the Royal Society.

If one neglects the finite width of the zone in the detonation front where chemical reactions occur, a freely running steady plane detonation front can only advance through an explosive with the Chapman-Jouguet velocity defined by $D = u + c$. Once the explosive products are formed, their subsequent chemical reactions and motion in the detonation front may be considered as adiabatic. Although Chapman (1899) and Jouguet (1901) correctly stated their equation, neither attempted to discuss the reaction zone itself. It was therefore thought necessary that the recent views on the reaction zone should be described in a manner which throws new light on the Chapman-Jouguet equation. Professor J. von Neumann, Dr S. F. Boys and

Dr A. F. Devonshire were the principal contributors on the theoretical side and von Neumann's theory (1942) will be outlined later.

The detonation velocity of the ideal plane detonation wave can be calculated if the equation of state of the products is known. Jones & Miller (1948) assumed an equation of state

$$pV/n = RT + bp + cp^2 + dp^3,$$

where p is the pressure, V is the specific volume, n is the number of moles of gas per unit volume, R is the gas constant and T is the absolute temperature. The value of n was fixed by making independent assumptions about the equilibrium constants for the various chemical reactions; and the values of the parameters b , c and d were adjusted to give agreement with the experimental values at $\rho = 1$ of D , $dD/d\rho$ and $d^2D/d\rho^2$, where ρ is the density of the explosive.

Kistiakowsky & Breit Wilson (1941) have made similar calculations using an equation of state of wider application

$$pV = nRT(1 + xe^{0.3x}),$$

where

$$x = K/VT^{\frac{1}{2}},$$

and K is about 12.5. This equation implies that the fugacities of all chemical species are increased in the same ratios at high densities, a similar assumption to that made by Jones & Miller except that they ignored the presence of ammonia.

Jones extended the theory to include the effect of the diameter of a cylindrical stick of explosive on the velocity of detonation and rate of reaction in the detonation zone. He has also discovered a method of finding an upper limit to the detonation pressure, independent of the assumed equation of state. This work is described later in the present discussion.

The motion of the gases behind a plane detonation front have recently been discussed by Paterson (1950), and comparisons made with drum camera photographs of the detonation of $C_2H_2 + O_2$, taken by Bone & Fraser (1931).

Sir Geoffrey Taylor (1950) first formulated correctly and solved the hydrodynamical equations in the steady plane detonation wave which begins at time zero at a rigid plane wall. An earlier attempt by Langweiler (1938), although containing results of interest, was not correct. Taylor obtained numerical solutions for various solid explosives using Jones & Miller's adiabatic equations for the explosive products, and also solved in similar fashion the problem of the steadily expanding spherical detonation wave which begins at a point at time zero. Numerical solutions for various solid explosives and gas mixtures were calculated.

Another 'similarity solution' of practical importance, solved numerically by Taylor & Jones (1942), was that finally reached by a plane detonation wave running parallel with the plane surface of a semi-infinite explosive, the space beyond the explosive being filled with a non-explosive fluid such as a perfect gas or water. An extension of this problem was solved numerically by Hill & Pack (1947). They solved, partly by numerical integration along the characteristics and partly by power series expansions in limited regions, the motion of the gases behind a detonating bare slab of explosive in air. The detonation zone was assumed normal to the two free surfaces and the motion was treated as irrotational. Both approximations are acceptable.

The theories of Jones, Kistiakowsky & Breit Wilson, Paterson and others on detonation velocities, and Taylor's theory of the plane and the spherically expanding detonation wave postulate a steady state, and where curved detonation surfaces are involved, treat the reaction zone as of negligible thickness. From the practical point of view it is of the greatest importance to understand how detonation starts and how long it takes for the detonation velocity to reach a steady condition. An account of early work is given in Marshall's *Explosives* (1932). Substantial advances have recently been made by Bowden, Courtney-Pratt and collaborators, as will be clear from a study of their papers presented in this discussion.

The detonation zone and the Chapman-Jouguet equation

The essential elements of a steady one-dimensional detonation process are illustrated in figure 1. If we succeed in understanding this process, various refinements can be introduced later.

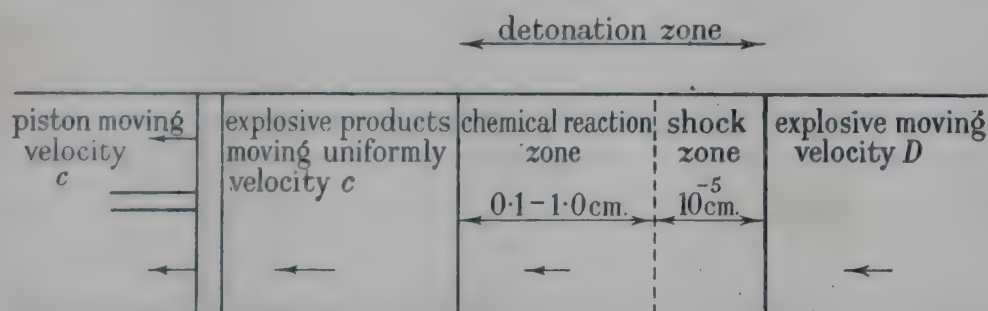


FIGURE 1. Steady plane detonation in a solid explosive. The detonation zone is reduced to rest by making the explosive feed into the detonation zone from right to left at the detonation velocity D .

An explosive material, solid, liquid or gaseous as the case may be, moves from right to left at the 'detonation velocity' D , and passes into the detonation zone which is stationary in space. Conditions throughout the detonation zone are steady. The material first passes into a plane shock zone where abrupt increases of pressure and temperature and a relatively small increase of density occur. The mass velocity of the material decreases as it moves through the shock zone to a value of order 50 to 80 % of D . The temperature throughout the shock zone, and indeed throughout the subsequent reaction zone, cannot be closely defined because the material is not in thermal equilibrium. The width of the shock zone is roughly the same as that of all shock zones, namely, 10^{-5} cm.

The tremendous increase in the thermal motion of the atoms or molecules as they pass into the shock zone causes chemical reactions to proceed with great rapidity. As the material moves to the left, reactions proceed, reach their limit and stop. The width of the reaction zone for military explosives is of order 1 mm. to 1 cm., but for other explosives, such as blasting explosives, it may be several times greater. An extreme case is that of an explosive dust mixture, such as coal dust suspended in air, where the reaction zone may be as great as 10 m. (Sir Geoffrey Taylor and others).

When the explosive products reach the end of the reaction zone, they are in chemical equilibrium; and the pressure, temperature, density, mass velocity and chemical composition are all uniquely determined. The explosive products can be

maintained at these conditions if they are contained by a piston which moves uniformly to the left at the same speed at which the products emerge from the reaction zone. This speed happens to be identical with the speed of sound in the gaseous products at these conditions, as we shall later demonstrate.

Steady and non-steady detonation

Suppose that at some stage where conditions are like those illustrated in figure 1, the velocity at which the piston is running away from the detonation zone is suddenly increased. The pressure in the explosive products near the piston will drop, and a rarefaction will spread back from the piston through the explosive products. The leading edge of this rarefaction will move with the speed of sound c relative to the products, but as the products themselves are moving with velocity c , the leading edge of the rarefaction will remain still in space, while the remainder of the rarefaction wave will be to the left of the leading edge, and will extend all the way to the piston.

An important limiting case is that where the piston starts from the end of the reaction zone and moves at constant velocity to the left with velocity D . We now have a plane detonation wave starting from a rigid wall, viewed from a system of reference in which the detonation wave is at rest. By the arguments already presented, the mass velocity of the explosive products as they emerge from the reaction zone is the local velocity of sound, and this result will still apply whatever the velocity of the piston, provided it is at least c . Let u be the mass velocity at the end of the reaction zone relative to the system of co-ordinates in which the undetonated explosive is at rest. Then we have

$$D = u + c.$$

This is the Chapman-Jouguet relationship.

Suppose now that conditions at some stage are like those shown in figure 1, and that the velocity of the piston is suddenly reduced to a constant value less than c . The explosive products start to pile up on the piston, and the result is a shock wave starting from the piston and moving through the explosive products. The velocity of this shock relative to the products is necessarily greater than c , and therefore the shock moves to the right and passes into the reaction zone and on to the shock zone. The detonation velocity then increases.

We see that in any freely running detonation, the detonation velocity must obey the Chapman-Jouguet condition, but that if the explosive products are forced forwards by a constraint which moves at a velocity greater than $(D - c)$, where D and c are the free-running detonation velocity and the velocity of sound at the end of the reaction zone in free detonation, the detonation velocity will increase beyond the free-running value. A conceivable method by which the detonation wave in an explosive can be 'overdriven' is to start the detonation by means of a more powerful explosive with a higher detonation velocity. The dynamical condition which must be fulfilled if one explosive is to be capable of overdriving another is easily formulated, but need not be given here. As an example of a pair of such explosives, however, we remark that if a stick of R.D.X./T.N.T. is placed end-on to a similar stick of T.N.T., and detonation proceeds from the former to the latter, one would expect for a short distance, at least, that the detonation in the T.N.T. would be faster than normal.

An interesting analogy exists between the regime of the plane steady detonation zone and the flow of a gas along a cylindrical tube whose walls contain a heater over part of their length. Extensive theoretical and experimental studies have been made of various cases of the latter problem. Reference may be made to the work of Hicks & Wasserman (1947).

The detonation zone

The detonation zone, as illustrated in figure 1, has been divided into two regions called the shock zone and the reaction zone. The division is helpful in visualizing the detonation process but is none the less somewhat artificial. It is realistic for a dust explosion because here the shock does heat up the gas molecules and the gas is roughly in thermal equilibrium before any appreciable burning of the dust particles occurs. Similarly, the division may be considered satisfactory for an explosive of the military type, partly because the reaction zone is very much wider than the shock zone and partly because conditions are in any case too complicated to permit more than a qualitative description. The division into two zones is not entirely satisfactory for a gaseous explosion, as will now be illustrated after a brief digression.

Experiments on the absorption and dispersion of ultrasonic waves in diatomic gases suggest that molecular vibrational energy is by no means as readily exchanged in collisions between molecules as is rotational or translational energy. Of course, these experiments do not permit the temperature of the gas to be high enough for collisions to be sufficiently energetic to knock the molecules into atoms. Again, experiments on the low-pressure flame spectra by Gaydon & Wolfhard (1949) have shown that the various degrees of freedom of certain diatomic radicals are not in thermal equilibrium. The evidence is consistent with the view that translational energy is very freely exchanged in a collision, but that rotational energy is not quite so readily exchanged. A typical 'hot molecule' with a rotational temperature of 9000°K appears to experience 40 collisions before coming into rotational equilibrium with its surroundings.

The above evidence suggests that it is reasonable to suppose that the shock zone in a gaseous detonation is at least roughly defined, and in this zone only translations and rotations have a high temperature. Subsequently, the other degrees of freedom are excited, and the chemical reactions proceed. If we accept this hypothesis, we can give some numerical values for any gaseous detonation. As an example, the observed detonation velocity in the $2\text{H}_2 + \text{O}_2$ detonation at atmospheric pressure is 2819 m./sec. (Lewis & von Elbe 1938). The pressure at the boundary of the reaction and shock zones is 32.8 atm. and the rotational-translational temperature is 1873°K . The mass velocity, relative to the undisturbed gas mixture, is 2282 m./sec. At the end of the reaction zone the pressure is 18 atm., the temperature is 3600°K and the mass velocity is 649 m./sec. The partial pressures in atmospheres of H_2O , H_2 , OH , H , O_2 and O are 10.33, 2.77, 2.25, 1.22, 0.83 and 0.60 respectively. These latter values, of course, can be computed without any reference to the reaction zone, as Lewis & von Elbe demonstrate.

The assumption made above that translations and rotations are excited and reach approximate thermal equilibrium before other degrees of freedom are affected may

not be justified. Should this be so, then the shock zone and the reaction zone become merged, and the width of the zone will be of order 10^{-4} to 10^{-5} cm.

An uncertainty similar to that just described for the detonation zone of a gaseous explosive is met in the shock zone of a powerful shock wave in air and other gases. For example, a shock of a few hundred atmospheres in air provides enough energy to lead to dissociation, excitation of vibrational and electronic states, and to the formation of oxides of nitrogen, but the details of the zone in which these reactions occur are too complicated to permit reasonable calculations to be made. On the other hand, conditions at exit from the zone, where thermal equilibrium has been reached, can quite simply be calculated, in close analogy with the similar calculations on gaseous detonation.

Von Neumann's description of the detonation process

Suppose that in the steady situation represented by figure 1, the pressure in the explosive before it enters the shock is p_0 and the volume of unit mass is V_0 . At some representative plane in the detonation zone, let V be the volume of unit mass and u be the mass velocity, defined in the manner used in the kinetic theory of gases. The fact that conditions are steady require that the pressure p is also defined. The temperature at the plane cannot be defined because thermal equilibrium does not obtain, but the partition of energy among the degrees of freedom of the various molecular and atomic species is uniquely determined, although varying from plane to plane. Hence the internal energy E per unit mass at each plane is determined, in spite of the absence of thermal equilibrium. Similarly, the amount of chemical energy Q which has been released per unit mass is defined.

The conservation equations of mass, momentum and energy are

$$\frac{u}{V} = \frac{D}{V_0} = m \quad (\text{a constant}), \quad (1)$$

$$p - p_0 = m(D - u), \quad (2)$$

$$E - E_0 = Q + \frac{1}{2}(D^2 - u^2) + p_0 V_0 - pV. \quad (3)$$

$$\text{From (1) and (2) follows} \quad p - p_0 = D^2(V_0 - V)/V_0^2. \quad (4)$$

Eliminating D from (3) and (4) gives

$$E - E_0 = Q + \frac{1}{2}(p + p_0)(V_0 - V). \quad (5)$$

Equations (4) and (5) can be represented by curves in the (p, V) plane. Thus, equation (4) gives a straight line of slope $-D^2/V_0^2$. As explained, for example, by Lewis & von Elbe (1938), the end-conditions of the reaction zone are defined by the point at which the line (4) is tangential to the particular curve (5) which passes through the end-conditions. Since this point represents a state of thermal equilibrium, it can be computed in the manner explained, for example, by Jones & Miller (1948). The helpful extension of the argument made by von Neumann (1942) was to sketch in other curves (5) corresponding with Q values lying between 0 and the value Q_f at the end of the reaction zone. It is reasonable to suppose that this family has the general form shown in figure 2.

The motion of a point on the (p, V) diagram representing the behaviour of an element of explosive passing into the detonation zone always remains on the straight line (4), because this relationship follows simply from continuity of mass and momentum. In the shock zone, the representative point runs from (p_0, V_0) up this straight line with great rapidity, reaching a limiting value at the point where it meets once again the curve (5) of zero Q . So far, we say that no chemical reactions have occurred, but the explosive material has been 'shocked' into a condition where reactions can proceed. The representative point now begins to return along the straight line, sweeping across curves (5) of gradually increasing Q values, until finally it reaches a Q curve to which it is tangential, and this is the Chapman-Jouguet point, where, once again, thermal equilibrium prevails.

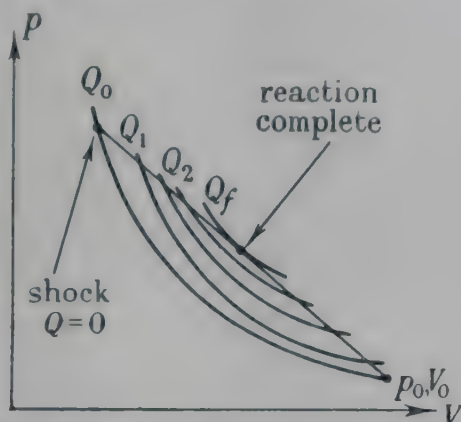


FIGURE 2. A diagram to explain von Neumann's presentation of the detonation process. The curves of varying Q values represent pressure-volume relationships of non-equilibrium states of partially reacted explosive. The curve $Q=0$ arises from a 'pure shock', while Q_f is the curve on which lies the point representing the final conditions in the reaction zone. As an element of material passes through the detonation zone, the point in the (p, V) diagram representing its behaviour moves only along the straight line, drawn through (p_0, V_0) .

Non-planar detonation fronts

Most real detonation fronts are curved, but the two radii of curvature are usually much greater than the width of the reaction zone. Curved detonation fronts may either be steady or non-steady. For example, the detonation wave passing along a long cylindrical stick of explosive, either cased or bare, presumably ultimately settles down to a steady state, although the front is not quite plane, but bends back a little at the edges because of the sideways expansion. The spherical detonation wave in an infinite explosive, originating from a small region at the centre, strictly speaking, reaches a steady state only asymptotically.

If L is the length of the reaction zone, R_1 and R_2 are the principal radii of curvature, and u and D are the mass velocity and detonation velocity for plane detonation, then the effect of the curvature of the front is of the same order as that produced by a change in loading density from ρ to

$$\rho[1 - uL/D(R_1 + R_2)].$$

For T.N.T. of loading density 1.5 g./cm.^3 , u/D is about 0.25 , and L is of order 0.2 cm. The effect on the detonation velocity and other variables of the spherical expansion of the detonation wave in T.N.T., when the wave has reached a radius 1 cm. , is there-

fore only of the order of that caused by a 2.5 % reduction in loading density, while the corresponding figure when the radius is 10 cm. is only 0.25 %. A 1 % drop of loading density causes the detonation velocity to fall by 65 m./sec., i.e. roughly 1 %; the drop in pressure at the end of the reaction zone is 1.7 %.

Courtney-Pratt (see F. P. Bowden's paper) has shown that the full detonation velocity is established in lead azide in a time less than $0.1 \mu\text{sec}$. This result implies that the reaction zone in lead azide is very narrow, possibly as small as 10^{-2} cm.

Jones (1947) has considered the manner in which the width of the reaction zone affects the velocity and shape of the detonation front in a detonating stick of explosive, while Paterson remarks in this discussion on the reduction of the reaction energy in cases of incomplete confinement. All authors agree that, provided detonation occurs, the errors caused by neglecting the width of the reaction zone in calculations on the mechanical variables of detonation are small.

There are, however, certain explosives which, when formed into thin sticks, often fail to detonate properly, even though the detonation is at first well established by means of a more reliable explosive. An example of such an explosive is T.N.T., in sticks of 1 in. diameter or less, cast in cardboard containers. Detonation of the first inch or two from the booster proceeds normally, but then the detonation sometimes stops, and the rest of the stick can be picked up relatively undamaged. Dr O. A. Gurton, in this discussion, describes some new experimental evidence on 'fading' in a coarse-grained explosive. The explanation of 'fading', as this effect is called, is undoubtedly dependent on the diameter of the stick, its confinement, the crystal size and the width of the reaction zone, but no satisfactory theory has yet been proposed.

SIMILARITY SOLUTIONS TO PROBLEMS INVOLVING GAS FLOW AND SHOCK WAVES

BY SIR GEOFFREY TAYLOR, F.R.S.

The most remarkable feature of detonation is the constancy of velocity of the detonation shock front which is found to have very nearly the same value whether the explosive is unconfined so that the burnt gases can escape freely, or confined in a tube which permits motion in one direction only. It is clear that the velocity of detonation must depend in some way on the mechanical reaction of the gases streaming away behind the detonation front. That this must be so can be seen by imagining a case when a rigid plane is forced to move forwards behind the front. If this plane were to move fast enough it must be able to increase the velocity of the detonation front beyond the value found when the burnt gas escapes freely.

The dynamics of the gas behind the detonation front must therefore be studied before an understanding of the mechanics of detonation can be attained. The general equations of the motions of compressible gases and of shock waves can be written down, but it has been possible to solve them only in special cases where the geometry of the flow field introduces elements of symmetry which enable great simplifications to be made. Instead of starting with general equations and then limiting their generality in successive stages in order to make possible another step forward in

analysis, I have often found it better to make the simplifying assumptions about symmetry at the beginning of the calculation. In such cases it is possible to guess the form of the solution and then subsequently prove that the guess was correct. Using methods of this kind it is possible to describe the field of flow behind a detonation front when detonation originates instantaneously at a point inside an explosive, so that the phenomenon has spherical symmetry; or in a tube so that the motion is one-dimensional (Taylor 1950*a*). It is found that in both cases the field of flow is similar at all times, but that its linear dimensions increase at a constant rate.

A list of problems in the flow of compressible fluids which can be solved owing to the simplicity of the geometry involved is given in table 1.

TABLE 1

	type of flow	dimensions in space	elements of symmetry
1	Prandtl-Meyer expansion round a corner (Prandtl 1907)	2	conditions constant along radii from corner (no shock waves)
2	radial flow outside a uniformly expanding sphere (Taylor 1946)	3	spherical symmetry conditions constant at corresponding points in a uniformly expanding frame
3	flow near wedge or cone in supersonic stream (Taylor & Maccoll 1933)	2 and 3	conditions constant at all times along straight lines through vertex
4	gas behind detonation front (Taylor 1950 <i>a</i>)	1, 2 or 3	conditions constant at corresponding points in uniformly expanding frame
5	detonation of explosive in a uniform tube which expands under the gas pressure (Taylor 1941)	3	the motion is axisymmetric and is constant relative to a frame which moves with the detonation velocity
6	blast wave from an intense explosion (Taylor 1950 <i>b</i>)	3	spherical symmetry: similarity of flow referred to axes which expand so that the linear scale $\propto t^{\frac{2}{3}}$: at similar points pressure $\propto t^{-\frac{2}{3}}$, velocity $\propto t^{-\frac{1}{3}}$ and density constant

THEORETICAL CONSIDERATIONS RELATING TO THE DETONATION OF SOLID EXPLOSIVES

BY H. JONES, *Imperial College, London*

Detonation in a cylindrical charge of solid explosive differs from detonation in a gas contained in a tube in several respects. Two important differences are: (*a*) the much higher temperatures and pressures which are developed in the detonation wave of a solid explosive and the greater complexity of the explosion products, and (*b*) the motion and the pressure gradients are not confined only to the direction of the axis of the solid charge on account of lateral expansion.

The modifications to the well-known Chapman-Jouguet theory required by (*a*) and (*b*) may be dealt with separately. If we imagine the charge to be so confined

that no lateral expansion is possible, the basic equations of the theory may be written

$$E = \frac{1}{2}p\left(\frac{1}{\rho} - v\right) + Q, \quad (1)$$

$$p = U^2\rho(1 - \rho v), \quad (2)$$

where E is the internal energy per unit mass of the explosion products, v their specific volume, Q the heat of the reaction and ρ the density of the unexploded charge. Equation (1) expresses the conservation of energy, the first term on the right denoting the work done by the pressure p of the detonation wave. Equation (2), in which U denotes the velocity of the detonation wave, follows from the equation of continuity and the conservation of momentum.

When E is expressed as a function of p and v (1) represents a curve in the p, v plane, and (2) a straight line. According to the Chapman-Jouguet theory the conditions in the detonation front are given by the values of p and v at the point where the straight line (2) touches the curve (1).

The greatest difficulty in applying this theory to solid explosives lies in the determination of E as a function of p and v . The main contribution to E is known to depend linearly on the temperature T , and in order to eliminate T it is necessary to know the equation of state for the explosion products. We have no direct knowledge of an equation of state applicable to gas mixtures at pressures of the order of 10^5 atm. and temperatures in the neighbourhood of 3000° K. In these circumstances recourse has been made in the past to the following procedure. A relation between p , v and T was assumed which contained a number of unspecified parameters. By comparing the calculated relation between U and ρ with the observed relation between these quantities the values of the parameters were determined. Applied in this sense, therefore, the theory provides estimates of the pressure, density and temperature in the detonation front.

An alternative procedure is as follows: We regard U as a known function of ρ and consider p , v and ρ as independent variables. Equations (1) and (2) then represent surfaces in the space p, v, ρ and the Chapman-Jouguet condition now requires that the surfaces shall touch along a certain curve. The normals to the two surfaces must have the same direction at each point of the line of contact and therefore

$$l/l' = m/m' = n/n', \quad (3)$$

where l, m, n and l', m', n' are the direction cosines of the normals to (1) and (2) at the points where the surfaces touch. The Chapman-Jouguet condition, in this form, yields therefore two independent equations. One of them leads to the usual theory already described; the other to the following general relation (Jones 1949):

$$p = U^2\rho/(2 + \alpha) \left\{ 1 + \frac{d \log U}{d \log \rho} \right\}, \quad (4)$$

where α is a necessarily positive quantity. Hence, letting $\alpha = 0$, this relation gives an exact upper limit to the detonation pressure when the observed relation between U and ρ is employed. This result is independent of any assumption regarding the equation of state.

The expression for α can be put in several equivalent forms; one which is convenient for calculating its value, although it does not make its essentially positive nature obvious, is the following:

$$\alpha = \left\{ \frac{C_p}{p} \left(\frac{\partial T}{\partial v} \right)_p - 1 \right\}^{-1}, \quad (5)$$

where C_p is the specific heat at constant pressure of the gas mixture.

If an approximate equation of state, obtained in the way indicated above, is used to calculate α , it is found that this number is not greater than about 0.2. It is this fact which makes equation (4) of some interest, for if, as a result of the uncertainty of the equation of state, α is in error to the extent of say 10 %, yet the value of the pressure in the detonation front will be accurate to 1 %.

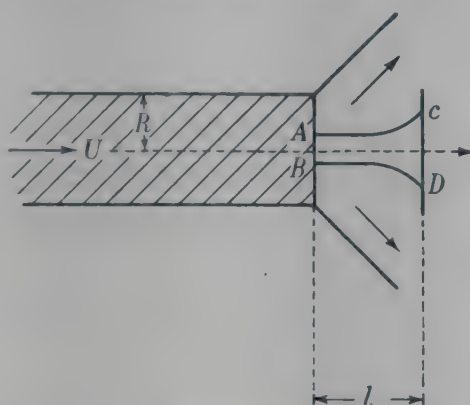


FIGURE 3

In order to determine the effect (b) of the lateral expansion on the detonation velocity it is convenient to regard the detonation wave as being at rest and the unexploded material as passing into the detonation front with the velocity U . It is sufficient, in the first place, to consider the conditions in a stream tube of small cross-section co-axial with the charge. If r denotes the relative expansion of the stream tube at the end of the reaction zone, which is of length l , i.e. if $r = CD/AB$ (see figure 3) and U_0 denotes the velocity of detonation when there is no lateral expansion, then it can be shown that

$$(U_0/U)^2 = 1 + c(r^4 - 1), \quad (6)$$

where c is a constant which is approximately equal to 2.0 for the explosive T.N.T. If R denotes the radius of the charge and l/R is small, as is generally the case under experimental conditions, $(r - 1)$ is small and (6) may be written

$$(U_0 - U)/U_0 = 2c(r - 1). \quad (7)$$

For any particular charge and conditions of confinement it is, in general, rather difficult to calculate r as a function of l . However, from similarity considerations it is clear that for bare charges r will depend upon l and R only through the ratio l/R . Even for cased charges this will remain true provided that the thickness of the casing remains in proportion to the radius of the charge. We can therefore write

$$r = f(l/R), \quad (8)$$

and

$$(U_0 - U)/U_0 = 2c\{f(l/R) - 1\}. \quad (9)$$

Now for a given charge R is fixed, therefore (8) is the equation of the curve AC of figure 3. On the other hand, for a series of charges of different R of the same explosive and therefore with the same l , we see that if $(U_0 - U)/U_0$ is plotted against $1/R$ we obtain again the curve AC , according to (9), except for a change of scale. The curve $(U_0 - U)/U_0$ against $1/R$ can, of course, be determined experimentally, and thus when the function f has been calculated a comparison between the scales of the observed curve and that given by (9) determines l and, therefore, the rate of the reaction in the detonation wave.

The general features of the $(U_0 - U)/U_0$, $1/R$ curves can readily be predicted without detailed calculations. In the case of a bare charge, assuming a plane detonation front, the curve AC of figure 3 must be horizontal at A and hence must lead to a curve of type (1), figure 4. However, the detonation front cannot, in general, be a plane perpendicular to the axis but must be curved in the way shown in figure 5.

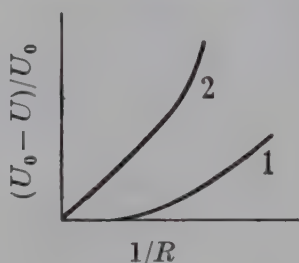


FIGURE 4

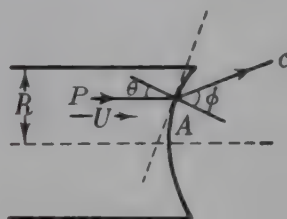


FIGURE 5

The reason for this is easy to see. A stream tube such as PAC (figure 5), which is parallel to the axis but some distance from it, will expand more than the central tube in the same distance l . Hence the velocity of the detonation front at this point, say U' , must be less than the velocity along the axis. Since the whole wave moves steadily with the velocity U it is clear that the front must be inclined at A as shown in figure 5, where $U' = U \cos \theta$.

The shock-wave equations applied at A show that the stream line AC is refracted away from the normal as shown in figure 5, i.e. $\phi > \theta$. Hence in reality, for a bare charge, the change in the velocity of detonation with $1/R$ will be as shown in curve (2), figure 4. If the radius of curvature of the detonation front at the axis, as well as the change in U with R , were determined for a single explosive it should be possible to determine the length of the reaction zone l with considerable certainty.

SOME CHARACTERISTICS OF DETONATION

BY W. M. EVANS, *Armament Research Establishment, Ministry of Supply*

[Plate 1]

Of the many physical quantities, such as pressure, temperature and velocity, involved in the complex phenomenon of the detonation of high explosives, one of the easiest and the most accurate to measure is the rate at which detonation is propagated in a charge, particularly when the charge is of simple geometrical shape

such as cylindrical or square section. In this short presentation measurements of velocity only are dealt with, and brief reference is made to the way velocity is related to various physical characteristics of the explosive charge.

Methods of measuring the velocity of propagation

Many methods have been evolved for measuring velocities. One that merits mention is that of Dautriche, which has given sterling service in the past and, indeed, for work of a routine nature is still much in use because of its simplicity and ease of operation (Copp & Ubbelohde 1948). Into the sides of the stick of explosive are inserted the two ends of a length of detonating fuse such as cordeaux or cordtex, the velocity v of which is known or can be taken as a standard. The ends are inserted at a distance x apart, and the middle region of the loop is straightened out and laid in firm contact with a lead plate. On detonation of the main charge, one end of the fuse is detonated at a time t before the other end, so that the two waves in the fuse meet in head-on collision at a point distance y from the mid-point of the fuse, the point of collision being denoted by an indentation on the plate. It follows that the velocity of detonation of the main charge is given by

$$V = \frac{vx}{2y}.$$

This method is, however, limited in precision and depends on constant propagation in the fuse, and for best results needs charges of the order of 20 in. in length.

A more elegant and accurate method is that in which the luminous trace of the detonation wave front is recorded photographically. The apparatus consists essentially of a lens assembly, a slit (either in the vicinity of the charge or in the lens system), a rotating mirror and a photographic plate or film. The light from the charge passes through the slit, the image of which falls on to the rotating mirror and then to a focus on the film (Cybulski, Payman & Woodhead 1949).

A mirror camera used in the Armaments Research Establishment to record phenomena such as are discussed later, is that built during the war by the Road Research Laboratory (1944). It is capable of recording luminous propagation in two directions at right angles, by means of two lens assemblies which have their optical axes in the vertical and horizontal planes containing the axis of the rotating mirror which is of square section, highly polished stainless steel driven at speeds up to 15,000 r.p.m. by means of a geared, low-voltage motor. Each lens assembly incorporates an adjustable slit (one vertical, one horizontal) which is focused on the film via a fixed reflecting mirror and the rotating mirror. The film is attached to a semi-circular film guide. Simultaneous vertical and horizontal luminosity traces are therefore recorded at a writing speed of up to 1 mm./ μ sec. with a maximum resolution of 0.2 μ sec.

The great advantage of the photographic method is that it gives a continuous record not only of the luminous propagation from its inception at initiation to its arrival at the end of the charge but the luminous shock wave propagated in the medium (usually air) beyond the charge. It can record the growth, the steady state and the dying-out of detonation. The greater part of the luminosity comes not from

the explosive wave itself but from the air shock in immediate association with the detonation wave. When the surrounding air is replaced by water or propane, for instance, the luminosity is greatly diminished, but when replaced by argon is very much increased. Many ingenious devices have been evolved for measuring the propagation in feebly luminous charges, in confined charges and of intense shocks through various materials.

Table 2 gives approximate values of the velocity of detonation of various explosives measured by various techniques. The velocity of detonation is, however, a complex function of the physical state of the explosive charge, and in any compilation of a table of accurate values the physical state of the charge such as density, grain size, confinement and diameter would need to be quoted. The few illustrations that follow are of obvious practical importance in design, but the relationships expressed are also of great fundamental importance in helping to elucidate the mechanism of propagation of detonation.

TABLE 2. APPROXIMATE VALUES OF VELOCITY OF DETONATION FOR
VARIOUS HIGH EXPLOSIVES

explosive	composition	density (g./cm. ³)	velocity of detonation (m./sec.)
picric acid	$C_6H_2(NO_2)_3OH$	1.6	7100
T.N.T.	$C_6H_2(NO_2)_3CH_3$	1.6	6900
tetryl	$C_6H_2(NO_2)_3N(CH_3)NO_2$	1.6	7500
P.E.T.N.	$C(CH_2ONO_2)_4$	1.6	7900
R.D.X.	$(CH_2NNO_2)_3$	1.6	8250
tetrytol 70/30	tetryl/T.N.T.	1.6	7250
amatol 50/50	ammonium nitrate/T.N.T.	1.6	5850
amatol 60/40	ammonium nitrate/T.N.T.	1.6	5600
amatol 80/20	ammonium nitrate/T.N.T.	1.6	5200
pentolite 25/75	P.E.T.N./T.N.T.	1.6	7000
cyclotol 50/50	R.D.X./T.N.T.	1.6	7600
torpex 42/40/18	R.D.X./T.N.T./aluminium	1.6	6800
baronal 50/35/15	barium nitrate/T.N.T./aluminium	2.3	5450

Variation of velocity with density

In general, the velocity is a linear function of the density of loading over a wide range of densities. For tetryl of crystal density 1.7 g./cm.³ the rate of increase of velocity with density is approximately 1000 m./sec. per 0.3 g./cm.³, a value which holds approximately for a large number of other explosives. It should be mentioned, however, that for a sufficiently low density propagation ceases altogether.

A probable reason for this is that as the density of packing is diminished the number of contacts between the grains or crystals is also diminished. Since it is these contacts that are subjected to the greatest compressional and frictional heating on passage of the wave, the number of 'hot spots' therefore decreases with decrease in density, and therefore propagation of reaction from grain to grain becomes less efficient until in the limit it finally ceases (Eyring, Powell, Duffey & Parlin 1949).

Variation of velocity with grain size

The velocity increases to an asymptotic value with decrease in grain size. If the grains or crystals are very large, as can be obtained, for instance, for T.N.T. when it is cooled slowly after casting in the mould, the charge fails to propagate and the well-known phenomenon of 'fading' occurs (Cybulski *et al.* 1949). Evidence has accumulated to show that on initiation by temperature each grain decomposes 'or burns' from the surface inwards and therefore takes a time, which can be called the reaction time, which is a function of the grain size. Therefore, the larger the grain, the longer the reaction time or the length of the reaction zone. When this length becomes of the same order as the radius of the charge, lateral losses by expansion play a very important role, for if such losses occur, the reaction can be considered as taking place at an effective lower loading density, so that the velocity of propagation is diminished and may die out altogether.

Variation of velocity with confinement

The velocity increases to a limiting value with increase in confinement of the explosive (Cybulski *et al.* 1949; Copp & Ubbelohde 1948).

Strength of confining tube has little or no influence, the main factors being the density (or its inertial effect) for a thin layer and compressibility for a thick layer. Thus a thick confinement of lead is not as good a substance as steel, for though it is of higher density its higher compressibility allows for greater expansion in the region of the reaction zone and thus leads to an effective lowering of the pressure in the reaction zone which is reflected in a lowering of the velocity of propagation.

Variation of velocity with diameter

The relationship between velocity of detonation and the diameter of the charge (Cybulski *et al.* 1949; Copp & Ubbelohde 1948), is of the same form as that exhibited by velocity and confinement. In fact, the two are equivalent, for the peripheral portions of a detonating charge can be considered to act as an effective and very high confinement to the central region. There is a limiting value to the velocity of an explosive. No further increase is obtained even if the charge is further increased in diameter or more heavily confined. A knowledge of this limiting velocity with that of the variation of velocity with diameter enables a value of the time of reaction of the explosive, or the length of the reaction zone to be calculated. For most granular explosives of high velocity this length is of the order of 1 mm. (10^{-7} sec.), and for mixtures of low velocity, such as T.N.T./ammonium nitrate, about 10 mm. (10^{-6} sec.). Direct experimental observation of reaction zone lengths has not yet been satisfactorily made: it is deduced indirectly from theory and experiment. Indeed, we are only at the beginning of our understanding of the mechanism of reaction in the wave, its propagation and the physical significance of factors such as the temperatures and the pressures involved.

Transition of detonation

Among many of the more recent investigations carried out at the Armaments Research Establishment may be mentioned one that deals with the transition of detonation from one explosive to another, in particular when the explosives differ

widely in velocity of detonation. For convenience cylindrical sticks are used, the two explosives being placed end to end in good contact. The R.R.L. camera has been used to record the progress of detonation in such composite charges.

Now experience of photographic and other techniques has shown that great care must be taken in interpreting the records of all but the very simplest of detonating systems. For a cylindrical stick of one component the record of the progress of stable detonation is usually straightforward of interpretation, but if, as is often done, the stick is encompassed by Cellophane or some other substance to increase its luminosity, or has various attachments, the external shock waves produced by the detonation may be reflected and reinforced and will travel forwards, highly luminous, and hence be recorded. For a composite charge as described above, the external peripheral shock wave produced travels forward with, and at the same speed as, the detonation wave in the first component, and on reaching the boundary between the

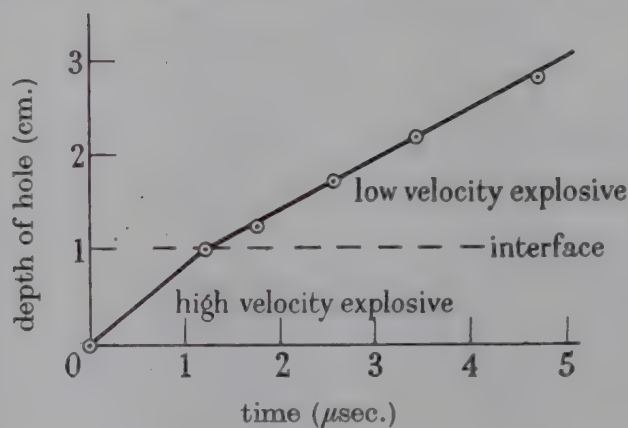


FIGURE 8

two explosives continues along the surface of the second component at a diminishing velocity which is higher for some distance than the velocity of detonation in that component. The record may well measure such enhanced external phenomena and may mask the record of the intrinsic propagation sought.

To avoid the possibility of measuring such spurious shocks the composite charges were arranged as follows: (a) Axial holes of various depths were drilled along a diameter, some passing through to the first component (figure 6, plate 1). The slit of the camera was focused across these holes so that the record obtained would represent the sequence in time of the appearance of luminosity at the bottom of each hole. (b) Holes perpendicular to the axis were drilled at regular intervals along the composite charge, and in each hole was inserted a tight-fitting copper tube which extended for half an inch or so above the surface of the charge (figure 7, plate 1). Plasticene was placed in the region between each tube and the slit focused as in (a) on the bottom of each hole.

The depth/time graph in figure 8 shows that in the region of the boundary between the two explosives a sudden drop in velocity occurs, that is, from about 8000 to about 5000 m./sec. This sudden transition is shown very clearly in figure 7, plate 1. For certain explosives, therefore, the velocity of the first component 'carries over' into the second component for a distance of certainly less than a few millimetres from the boundary. The same effect is shown when the transition is from the low component

time of appearance of
light at bottom of holes



position along diameter



photograph of drilled charges

FIGURE 6. Method of determining 'carry over' distance

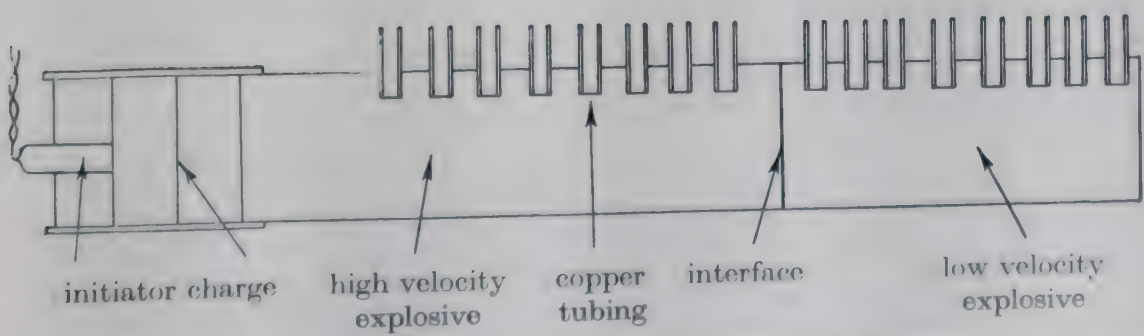
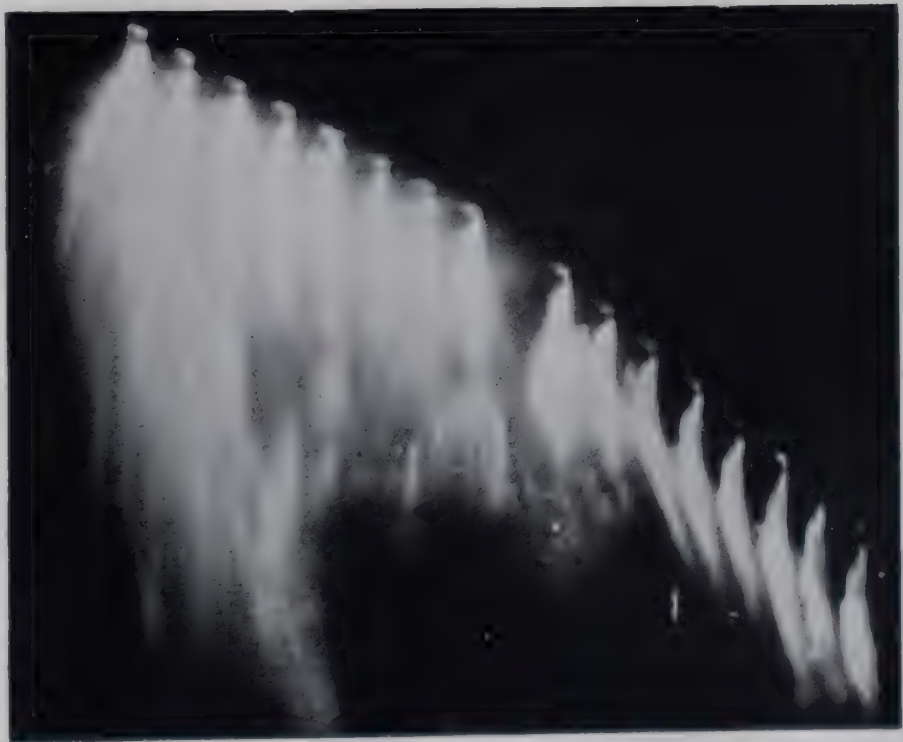
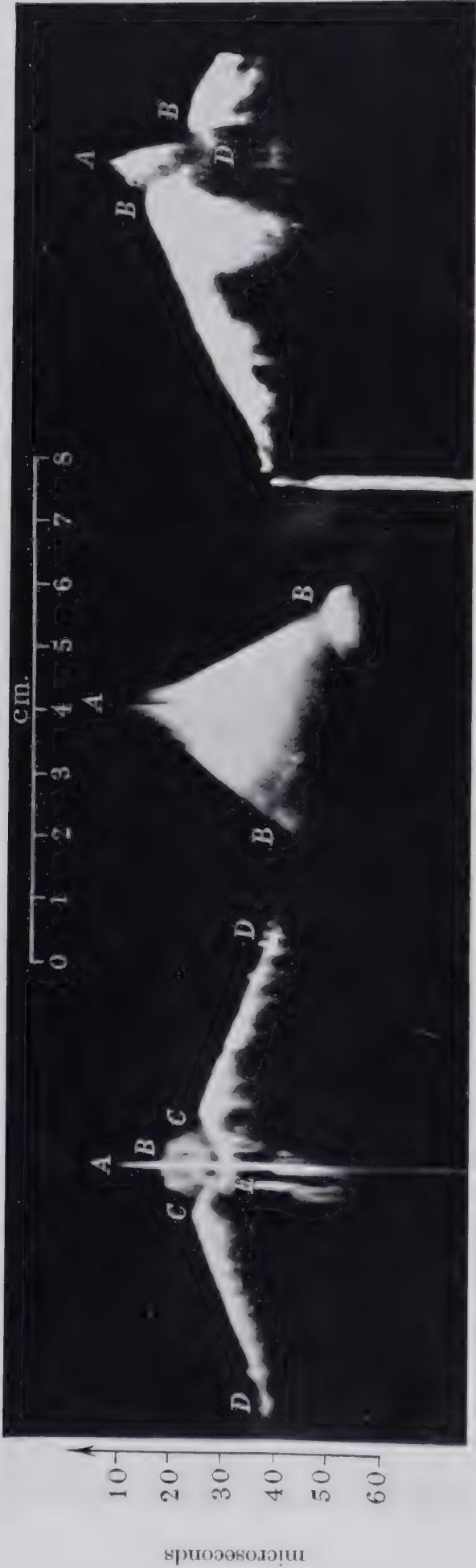


FIGURE 7. Method of determining 'carry over' distance.



a

FIGURE 9a. Explosion of a confined film of nitroglycerine by the impact of a cavity striker. Initiation occurs inside the cavity at *A*, and the slow-burning *AB* (*c.* 10 m. sec.) is followed by a rapid burning in the confined film *BC*. Detonation *CD* at about 220 m. sec. is the next stage in the explosion. Retonation waves *CE* travel back from the points at which detonation is first observed.

b

scale of 1 cm. (measured in the plane of the explosion)



c



FIGURE 9b. Explosion of a ring of nitroglycerine set off by impact. The point of initiation *A* corresponds to a compressed air bubble, and the explosion *AB* spreads as a rapid accelerating burning at 180 to 650 m. sec.

FIGURE 9c. Explosion of P.E.T.N. initiated by impact on a ring of explosive propagated into a surrounding layer of confined but unimpacted explosive. In the impacted area only the rapid-burning stage *AB* is observed, but the surrounding explosive detonates *BC*. Retonation waves *BD* travel back into the burnt gas of the first stage of explosion.

FIGURE 9d. Explosion trace for lead azide, photographed by Courtney-Pratt's camera. Detonation at *c.* 2000 m. sec.; no burning stage is observed. This photograph is reproduced five-eighths of the size that it appeared on the fluorescent screen, and shows the progress of the detonation over a distance of 4 mm. on either side of the point of initiation.

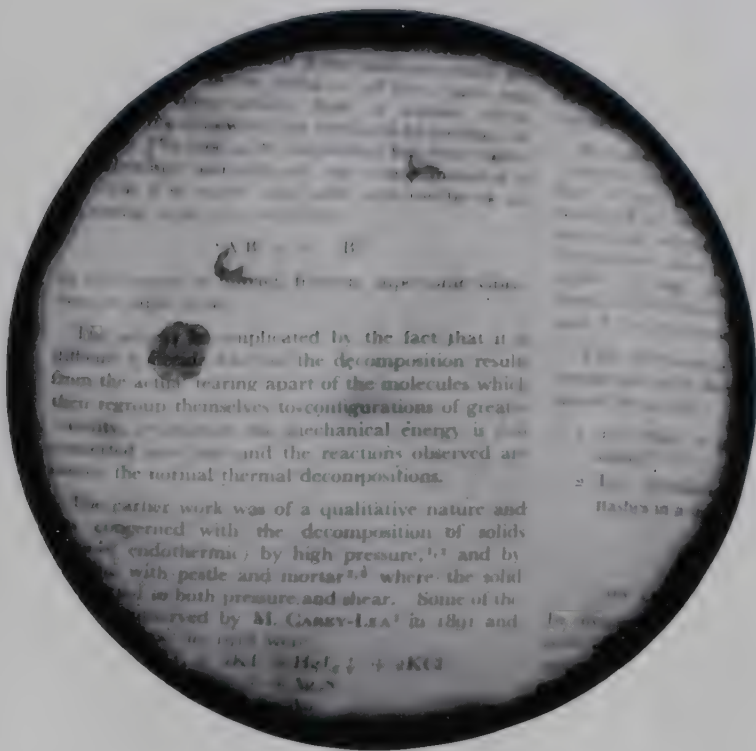


FIGURE 11. A photograph of the fluorescent screen of the image converter tube showing the image of a page of print transmitted through the tube.

motion of film



FIGURE 13. Fading to failure in a cartridge of T.N.T. Note the tendency for the velocity to remain constant first at high velocity (3800 m./sec.) and then at low velocity (1800 m./sec.).

to the high. This result is, however, not obtained if, for instance, the second component is loosely packed and therefore of low density. In this case the transition in velocity is gradual and takes several centimetres to reach a characteristic steady velocity. In fact, the abrupt transition depicted above appears to be true only for certain explosive combinations, the gradual transition being the more general case (Eyring *et al.* 1949; Cybulski *et al.* 1949). The mechanism by which transition occurs for the many explosive combinations examined is not yet fully understood.

PRESSURE MEASUREMENTS IN DETONATING GASES

(a) *A new pressure-bar technique*

BY R. M. DAVIES AND J. D. OWEN

Department of Physics, University College of Wales, Aberystwyth

The Hopkinson pressure-bar, in the original form (Hopkinson 1914) and in the electrical version (Davies 1948), suffers from the inherent theoretical limitation that the longitudinal elastic waves propagated in the bar show dispersion, and, as a result, stress pulses travelling along the bar suffer distortion. The dispersion and the associated distortion have been investigated experimentally, and the results have been found to be in excellent agreement with the accurate theory of the longitudinal vibrations of a circular cylinder due to Pochhammer (1876) and to Chree (1889) (see Love 1934, §201). It is interesting to note that, according to this theory, torsional waves in a circular cylinder do not suffer dispersion (Love 1934, §200), provided that the angular displacement is a linear function of distance measured from the axis of the cylinder, i.e. provided that the higher modes of vibration are not excited.

It therefore seemed worth while to design a pressure-bar in which the applied pressure was used to excite torsional waves rather than longitudinal waves, and a brief account of an experimental technique suitable for this purpose has been published recently (Owen & Davies 1949). The angular displacement of a small, optically polished flat on the surface of the bar, due to the torsional waves, is recorded using a high-speed revolving mirror which gives a writing speed of order 4 mm./ μ sec. on the film. The preliminary experimental results are promising; for example, the analysis of a record of the pressure caused by the impact of a flat-headed lead bullet shows that the initial, sudden pressure rise from zero occurs in less than 1 μ sec., compared with 5 μ sec., or more, observed with an ordinary pressure-bar. The method is now being developed to measure pressures in detonating gases.

(b) *Preliminary results on measurements of pressures in detonating gases*

BY R. M. DAVIES, D. H. EDWARDS AND D. E. THOMAS

Department of Physics, University College of Wales, Aberystwyth

During the war, Sir Geoffrey Taylor suggested that it would be interesting to use the electrical modification of the Hopkinson pressure-bar to measure pressures in detonating gases, and some preliminary experiments gave promising results. This

work has recently been continued, using gaseous mixtures which are detonated in a steel explosion tube, 12 ft. long and $1\frac{1}{4}$ in. square section. The pressure gauges can be set in two positions relative to the detonation wave. In the first position, the gauge is placed with its sensitive face flush with the side-wall of the explosion tube; the quantity then measured is the 'static' pressure in the wave, i.e. the discontinuous jump in pressure which occurs when the wave travels past the gauge. In the second position, the gauge is mounted in the closed end of the tube, and when its face is flush with the end, it measured the 'reflexion' pressure, i.e. the static pressure, together with the dynamic pressure due to reflexion at the closed end.

Three types of gauges have been used: (1) quartz piezo-electric gauges, (2) an electrical pressure-bar using a parallel-plate condenser unit (Davies 1948), and (3) a pressure-bar in which a resistor strain gauge is used to measure the longitudinal elastic strain on the cylindrical surface of the bar.

The chief difficulty in the measurement of these pressures arises from the extremely sudden initial rise of pressure. In the quartz gauges this excites the natural modes of vibration, and all records show pronounced oscillations following the initial rise, which is completed in 2 or 3 μ sec. The oscillations, however, occur around a level which is sensibly constant for times of order 100 μ sec., and in calculating pressures from the oscillograms, the average pressure, deduced from the measurement of the area under the curve, has been taken as correct. The scatter of the experimental results can be judged from the results of nine independent determinations of the excess reflexion pressure in the mixture H_2 , $\frac{1}{2}O_2$: 34.6, 32.4, 35.3, 33.4, 33.9, 33.8, 30.1, 33.0 and 35.8 atm., giving a mean value of 33.6 atm.

The uncertainty caused by gauge oscillations can be avoided by using a pressure-bar, and, although this type of experiment is subject to errors caused by dispersion, the bar methods are, on the whole, more accurate than those employing piezo-electric elements, in cases where a sudden discontinuity in pressure occurs. One disadvantage of the electrical pressure-bar used in conjunction with a parallel-plate condenser unit arises from the necessity of differentiating the oscillogram in order to obtain the (pressure, time) curve, and it was to avoid this difficulty that the experiments with resistor strain gauges were undertaken. These gauges, which were cemented so that their lengths were parallel to the axis of the bar, measured the longitudinal strain on the free surface of the bar. There are two disadvantages attached to this method of using a pressure-bar; on the one hand, the dispersion effects in a bar are known to be most pronounced on the free surface, and, on the other, the finite length of the gauges implies that they will not resolve variations in strain which occur in times which are of the same order of magnitude as the time taken by a stress wave to traverse the length of the gauge (1 μ sec. for the gauges of length $\frac{1}{4}$ in. used in these experiments).

On the whole there is little to choose between these two methods of using a pressure-bar, particularly when, as in the present instance, the pressure remains sensibly constant after a sudden initial jump.

The experimental results which have been obtained by the three methods are summarized in table 3. In the second column, *R* and *S* indicate measurements of the reflexion and static pressures respectively. The third, fourth and fifth columns give

the mean values of the excess pressures, expressed in atmospheres, the number of independent experiments corresponding to each result being quoted in parentheses. The calculated values of the static pressures given in the sixth column are those given by Jouguet (1905-6), who assumed that the specific heat of the detonating gases at constant volume was a linear function of the absolute temperature; the values quoted for the reflexion pressure have been calculated from the data used by Jouguet, using the theory given by Sir Geoffrey Taylor in 1940 in an unpublished report to the Ministry of Home Security as a basis, and extending it to the case of a gas whose specific heat varies with temperature according to the law assumed by Jouguet.

TABLE 3

		measured values (atm.)			
			pressure-bar		
mixture		quartz gauge	resistor gauge	condenser	calculated values (atm.)
C ₂ H ₂ , O ₂	R	—	77.0 (1)	69.0 (3)	269
C ₂ H ₂ , 2½O ₂	R	—	57.2 (2)	—	—
H ₂ , ½O ₂	R	33.6 (9)	35.9 (1)	32.6 (1)	84.7
C ₂ H ₂ , O ₂	S	47.8 (3)	40.1 (2)	—	53.5
C ₂ H ₂ , 2½O ₂ .	S	37.7 (4)	—	—	33.0
H ₂ , ½O ₂	S	18.2 (3)	—	—	16.5

TWO NEW STREAK CAMERAS

By C. A. ADAMS, *Armament Research Establishment, Ministry of Supply*

An optical streak camera designed for observation of explosions on the field trial scale has recently been developed in the Armament Research Establishment. The camera uses a two-lens system with the slit in the intermediate image position, as in the Road Research camera described by Dr Evans. It incorporates two optical systems for simultaneous observation of a horizontal and a vertical slit, and two independent film quadrants. Since it is designed to give a high writing speed, obtained from a $2\frac{1}{2}$ in. square-section mirror block rotating at 40,000 r.p.m. and an effective track radius of about 19 in., and employs large objectives for viewing distant events, a heavy and fairly bulky assembly is necessary. It is carried on a mounting ring which allows 90° rotation and is supported on a trolley to which are fixed the control panel and a pump for partial evacuation of the rotating mirror chamber. Provision is made for interchanging objective lenses to cover a range of distances from 20 ft. to infinity. For distances of the order of 100 yards from the event the objectives are 36 in. telephoto lenses. Cassettes are supplied for daylight loading. The film width is 90 mm. (3.54 in.) of which $2\frac{3}{4}$ in. is available for recording, the remaining $\frac{3}{8}$ in. on each side being required for support at the edges. Using film of 0.008 in. thickness, with controlled tensioning, errors due to film distortion are believed to be unimportant. A small error is caused by the divergence of the image path from a circle, but this error has been minimized by fitting each of the quadrants which support the film to the 'best' circular quadrant over the image path. In addition to adjustments

for orientation a focusing provision is made for lateral shift of the objectives to ensure that the system is used at maximum aperture.

Arrangements for speed measurement and synchronization are based on a magnetic pick-up from the mirror shaft. The output is fed into a tachometer system which shows the speed of rotation on two meters, one coarse and the other fine. The firing impulse is taken from the same circuit, and hence can be synchronized to any required phase of the mirror. The priming operation to render this circuit operative is taken from the mechanical shutters included in the system. The writing speed for a rotational speed of 40,000 r.p.m. is approximately 4 mm./ μ sec.

A second streak camera, based on an air-driven motor, has been constructed and used. The rotational speed obtained in this case is 80,000 r.p.m., giving a writing speed of 6.5 mm./ μ sec. A laminar mirror approximately $1\frac{1}{2} \times 1 \times 0.1$ in. is mounted on the motor shaft, and hence the advantage of higher speed is offset by the limitation to a single-slit system and by some reduction of aperture. The tachometer and synchronizer used with this camera are identical in principle with those used in the electrically driven camera, but another timing device has been added. This consists of three spark gaps mounted inside the main camera body, with a triggering circuit which ensures that the discharge in each gap shall occur during the recording time. The spark images are brought on to the recording film, at a position displaced from the central image path. The instants of discharge are monitored on an oscillograph, and hence the time scale and origin of the streak record are independently determined.

THE INITIATION OF AN EXPLOSION AND ITS GROWTH TO DETONATION

BY F. P. BOWDEN, F.R.S.

*Research Laboratory on the Physics and Chemistry of Rubbing Solids,
Department of Physical Chemistry, University of Cambridge*

The contributions we have just heard have dealt with the propagation of a stable detonation through a mass of the explosive. The velocity of this with many explosives is high, as Dr Evans's tables illustrate it may be of the order of 5000 to 8000 m./sec. Dr Penney, Sir Geoffrey Taylor and Professor Jones have shown this stable high-velocity detonation can be explained on fairly simple physical concepts. I would like to carry the discussion back a stage further—or perhaps two stages further—to the initiation and to the growth of the tiny explosion in liquids and solids *before* it has grown to a full-scale high-speed detonation.

First let us consider in a very elementary way the cause of initiation by impact or by friction. How is the mechanical energy of the blow or of rubbing able to produce the chemical excitation which results in reaction? There is evidence that the chemical reaction is not initiated directly by the shock nor by direct rupture of the molecules nor by rapid shearing or adjacent molecular layers. Apparently the mechanical energy of the blow must first be concentrated and degraded into heat to form a localized 'hot spot' of small but finite size. Thermal ignition of the

explosive then occurs at this hot spot. The diameter of these local hot spots may perhaps be of the order of 10^{-5} to 10^{-3} cm.

We have found that the hot spots can be formed in three main ways:

- (1) by the adiabatic compression of small entrapped bubbles or gas spaces,
- (2) by frictional hot spots formed
 - (a) on the confining surfaces,
 - (b) on extraneous grit particles,
 - (c) by intercrystalline friction of the explosive particles themselves,
- (3) by viscous heating of the rapidly flowing explosive as it escapes from between the impacting surfaces.

Both theory and experiment suggest that the third method is operative only under extreme conditions. It is the first two which are of the greatest significance. We may demonstrate this by a few simple experiments.

Initiation by adiabatic compression of included gas

If we place a film of a liquid explosive such as nitroglycerine on an anvil and hit it with a flat hammer we find that comparatively high impact energies are necessary to initiate the explosion. With a simple fall hammer, for example, an impact energy of 60,000 g.cm. (1000 g. falling 60 cm.) will give an 'explosion efficiency' of 45 %. If, however, a small pin hole or cavity is made in the anvil or striker, or if the explosive is initially distributed on the anvil as droplets or as an annulus or as parallel strips (instead of as a continuous film), there is an enormous increase in the impact sensitivity. If, for example, the striker contains a small cavity about 0.5 mm. in diameter, the explosion efficiency is 100 % with a mass of 40 g. falling 10 cm. (energy 400 g.cm.). Occasional explosions may even occur when the height of fall is only 0.5 cm. (energy 20 g.cm.) (Bowden, Mulcahy, Vines & Yoffe 1947).

The sensitivity is due to the fact that very small air bubbles are entrapped during impact, are heated by adiabatic compression, and fire the explosive. These results are quite general and experiments show that they apply not only to most liquid, gelatinous, and plastic explosives but also to solid explosives as well. The gas entrapped between the crystals and included in the solid explosive acts in a similar manner. The size of the gas pockets necessary to initiate the explosion may be very small, perhaps less than 10^{-3} cm. in diameter, so that they may readily escape detection.

Since the temperature of the compressed gas is given by $T_2 = T_1 \left(\frac{p_2}{p_1} \right)^{\left(\frac{\gamma-1}{\gamma} \right)}$ it depends mainly upon the physical nature of the gas and on the compression ratio. There is evidence that a momentary temperature rise in the tiny bubble of the order of 400 to 500° C will bring about the initiation of a wide range of explosives. A temperature rise of this magnitude would, with air, be produced by a compression ratio of about 30:1, so that it can, with a tiny bubble, be produced by a very gentle blow.

Frictional hot spots

When any two solids are rubbed together local hot spots are readily formed on the summits of the surface irregularities. With metals we may measure these

momentary hot spots by using two different metals as a thermocouple, sliding them together, and measuring the thermal p.d. developed. With non-conductors such as glass, quartz or transparent crystals we may observe the hot spots visually, or may detect them by lead sulphide cells which are sensitive to flashes of infra-red radiation. These methods all give similar results and show that local temperature flashes, of 1000°C or more, may easily occur on the surfaces, even under comparatively gentle conditions of sliding. The temperature rise is naturally limited by the melting-point of the solid and, in general, does not rise above it.

Measurements have been made, by the thermoelectric method, of the surface temperature of rubbing metals in the presence of liquid explosives such as nitroglycerine or nitroglycerol. At the same time the incidence of explosion has been recorded photoelectrically (Bowden, Stone & Tudor 1947). It is found that ignition of the explosive can occur when the temperature of the surface hot spots (as recorded by the thermoelectric measurements) reaches a value of *c.* 450°C or more. At temperatures below this initiation is not observed.

In a second series of experiments these same explosives were rubbed between glass and various metals. The metals or metallic alloys were selected with a range of melting-points, since, as we have seen, this forms a convenient method of limiting the temperature of the hot spot. It was found that with metals or metallic alloys melting below 450°C no explosion occurred even at very high loads and speeds. An alloy melting at 480°C and all metals melting at higher temperatures, however, cause explosions. The incidence of explosion was not determined by the hardness of the metal but simply by its melting-point.

It is well known that the presence of grit particles will render many explosives more sensitive both to impact and to friction. This provides an interesting method of introducing into the explosive hot spots of known maximum temperature, since the hot spot on the particle will, in general, be limited by its melting-point. 'Grit' particles of known melting-point were therefore added to a range of solid and liquid explosives, and their influence on impact and friction sensitivity was studied. The 'grit' particles were of widely different chemical composition but with similar hardness. Some typical results (Bowden & Gurton 1949), for the solid secondary explosive P.E.T.N. are given in table 4.

These results are very striking. There is no obvious correlation between the sensitivity and the hardness or chemical properties of the grit, but if it is compared with the melting-point (col. 3) a remarkably sharp division is apparent. All grits of melting-points greater than 430°C are effective in causing explosion while all grits of melting-point less than 400°C are ineffective. Other secondary explosions, whether solid (e.g. cyclonite) or liquid (e.g. nitroglycerine), behaved in a similar manner.

The primary explosives, such as lead azide, lead styphnate, mercury fulminate, also gave similar results, although the limiting melting-point of the grit varied with the nature of the explosive and was somewhat higher (500 to 550°C).

It was observed that when the size of the grit particle was decreased below a few micra its sensitizing effect diminished, but Mr Williams has recently found that particles of carborundum or of diamond as small as 5×10^{-5} cm. in diameter still

TABLE 4. INITIATION OF P.E.T.N. BY FRICTION AND BY IMPACT IN THE PRESENCE OF GRIT OF VARYING MELTING-POINTS

grit added	hardness (Mohs)	m.p. (° C)	explosion efficiency (friction) %	explosion efficiency (impact) %
nil	—	—	0	2
ammonium nitrate	2 to 3	169.6	0	2.5
potassium bisulphate	3	210	0	2.5
silver nitrate	2 to 3	212	0	2
sodium dichromate	2 to 3	320	0	0
sodium acetate	1 to 5	324	0	0
potassium nitrate	2 to 3	334	0	0
potassium dichromate	2 to 3	398	0	0
silver bromide	2 to 3	434	50	6
lead chloride	2 to 3	501	60	27
silver iodide	2 to 3	550	100	—
borax	3 to 4	560	100	30
bismuthinite	2 to 2.5	685	100	42
glass	7	800	100	100
rock salt	2 to 2.5	804	50	6
chalcocite	3 to 3.5	1100	100	50
galena	2.5 to 2.7	1114	100	60
calcite	3	1339	100	43

have an appreciable sensitizing influence. It is probable that most explosives even when very carefully prepared will contain foreign particles which are larger than this.

An interesting difference may be noted between the primary and the secondary explosives. Most of the secondary explosives melt at a temperature *below* that at which explosive decomposition sets in. Intergranular friction between the crystals of the explosive will not in general cause initiation because local melting occurs at the points of rubbing contact. With the majority of the primary explosives, however, explosive decomposition occurs below the melting-point, so that inter-crystalline friction may initiate the explosion.

Initiation in the gas phase

There is evidence that the initiation of the explosion begins in the vapour phase. When initiation is brought about by adiabatic compression of a tiny bubble, it would seem that the reaction begins in the hot gas inside the bubble and is then communicated to the condensed explosive. It is probable that this vapour phase initiation may also occur when the hot spots are formed in other ways, e.g. by friction or by sparking.

Some recent experiments by Gray & Yoffe (1950) have shown that ignition can occur when the concentration of the explosive vapour is remarkably low. With methyl nitrate vapour, for example, it is found that inflammation can occur at a temperature of c. 400° C when the concentration of the explosive vapour is as low as 1 part in a million. The reaction begins as a 'blue glow' which precedes the

normal explosion. Their results suggest that the explosion propagates through the vapour, not by a thermal mechanism as has been suggested, but by a radical chain or chain thermal mechanism.

Growth to detonation

A photographic investigation of the growth of the explosion in the immediate vicinity of the point of initiation brings out a number of points of general interest. The first is that frequently it does not grow at all: it is stillborn. Although there may be no outward sign of an explosion, sensitive photographic and other methods show that a micro-explosion has occurred and has died away without propagating to an appreciable distance. This observation that the small explosions can occur without our being aware of them is of some practical interest.

Moving-film studies show that, in addition to being a precarious process, the development of the explosion during the first few microseconds after its birth or during the first few millimetres of its growth is a complex one.

Figure 9a, plate 2, shows a moving-film picture of the growth of an explosion in a confined film of nictroglycerine initiated by the impact of a cavity striker. Initiation begins as a slow burning (*c.* 10 m./sec.) inside the cavity at *A* followed by a more rapid and *accelerating* burning in the confined film *BC*, a *constant-velocity low-speed* detonation (*c.* 2000 m./sec.) then sets in *CD*. In this photograph the rapid burning stage is somewhat obscured by the geometry of the apparatus, but it is shown more clearly in figure 9b.

Most solid explosives behave in a similar manner (figure 9c). There is again a preliminary burning which precedes the low-velocity detonation of *c.* 2000 m./sec. An exception to this is lead azide. Here the preliminary burning stage is absent and detonation sets in immediately (see figure 9d). This photograph was taken on Dr Courtney-Pratt's image converter camera (see below). This single-stage decomposition may be associated with the fact that the azide molecule is a simple one which decomposes only in one way.

It is probable that the speed of burning in the first stage represents a mass movement of the gas products away from the point of initiation. The second stage of constant-velocity detonation may be identified with the *low-velocity* detonation observed in large charges.

In the high-velocity detonation the shock wave is sufficiently intense to raise the temperature of the condensed explosive to a high value ($> 1000^{\circ}\text{C}$) by adiabatic compression. In the low-velocity detonation, however, the shock wave could not, by this means, raise the temperature more than *c.* 40°C . How then can the low-velocity propagation continue? We have suggested that two mechanisms can be operative here. With mobile liquids the shock front can facilitate propagation by breaking up the liquid into small droplets which are thrown into the reaction zone and present a large surface so that rapid reaction can occur. With solids and with viscous liquids, it is suggested that the formation of local hot spots play an important part. Once again the commonest sources of hot spots are small gas pockets which are present in the explosive and are suddenly compressed by the shock front. This heated interstitial gas then ignites the explosive. With the primary explosives hot spots can be formed by inter-crystalline friction. Under appropriate

conditions, therefore, *local hot spots may be just as important in propagation as they are in initiation.*

The chemical reactions associated with the different physical stages of the explosion are themselves different. Robertson & Yoffe (1948) have analyzed the decomposition products formed during the early stage of the explosion, and shown that they correspond more closely to the thermal decomposition products than they do those formed by complete high-velocity detonation. It is clear therefore that the early stages of the explosion involve chemical processes and physical processes which are complex and are different from those operating during the high-speed stable detonation. Further work will have to be done before we can say exactly how and when a deflagration will pass over into a detonation, or a low-velocity detonation step up to a high one.

THE PHYSICAL CHEMISTRY OF DETONATIVE REACTIONS

BY A. R. UBBELOHDE, *Queen's University, Belfast*

The Chapman-Hugoniot equations relate the velocity of a detonation wave in a system, to the energy release ΔE , and to variables such as the compression ratio and the specific heats of the system. No reference is made to the *mechanism* of the physico-chemical transformations in the detonation wave which result in the energy release.

Three points may be made about the physical chemistry of possible detonative reactions in a system.

(a) *Experimental evaluation of the temperature coefficient for the reactions leading to energy release*

Experiments have been made on the effect of the temperature of the reaction zone, on the rate of energy release of an explosive such as T.N.T., by measuring detonation rates of mixtures of this explosive with various amounts of a compressible inert component such as NaCl. From the effects of charge diameter on detonation rate of T.N.T./salt mixtures, it can be shown that a lowering of temperature of the reaction zone by several hundred degrees does not lengthen the time of energy release by more than a small factor (Copp & Ubbelohde 1948).

It follows that the chemical transformations in the detonation of T.N.T. do not depend on selective Arrhenius activation processes. The reason for this can be readily understood from a consideration of the high concentration of free radicals and atoms in the detonation zone, and of the effect of mass motion in contributing to the activation (Ubbelohde 1949).

(b) *Conditions for detonative transformation in physico-chemical systems with small ΔE —possibility of 'weak detonation' waves*

With the military high explosives, ΔE per unit volume is large, and the requisite conditions for activating the physico-chemical transformations can be attained fairly readily, as the temperatures in the reaction zones are high.

A less familiar problem is, whether there are any fundamental limitations to 'detonative transformations' at the appropriate Chapman-Hugoniot rate in physico-chemical systems for which ΔE may be quite small; for example, chemical reactions with only small heat evolution, or even polymorphic transformations. An illustration is the case of arsenic, which transforms explosively to a crystalline polymorph (Geiling & Richter 1949).

For reactions with small ΔE , the temperature rise in the detonation would tend to be small likewise, and the rate of energy release would be slow. This would involve very broad reaction zones. Unless the detonating system is very strongly confined, or unless the detonation is taking place over very large areas (effectively, with a very large value of the charge diameter R), the lateral energy losses would lead to quenching of such weak detonation waves. Very tentatively, and following the principle of similitude, a limiting condition for weak detonation waves may be suggested as follows:

Suppose stable detonation can be observed for bare charges of diameter R_1 in a system (high explosive) where the reaction zone has a thickness X_1 . In a system where the (weak) detonation would require a reaction zone of thickness X_2 to complete the energy release, stable detonation would only be observable in a charge diameter $\geq R_2$, where

$$R_2 = R_1 X_2 / X_1. \quad (1)$$

The establishment of detonative transformations in systems with small ΔE and large X_2 would thus require very large diameters. One example of this appears to be NH_4NO_3 . It is known that very large masses of NH_4NO_3 can be brought to detonation, as in the Oppau disaster, and in certain shipping accidents. Although the nitrate probably requires to be sensitized by combustible impurities to show this effect, it seems likewise important from the above considerations that the chemical decomposition must be set in train over comparatively very large areas if detonation of such weak explosives is to be set up. This has a bearing on safety precautions.

(c) *Possible geophysical examples of 'weak' detonation waves*

If the tentative similitude relationship (1) is well-founded, the only reason why the majority of familiar exothermic physico-chemical processes are not observed to occur at the Chapman-Hugoniot rate is that normally they are not studied on a sufficiently large scale. Such large-scale transformations may, however, be important in the earth's crust. Following the rule of successive states (cf. Ubbelohde 1937) the solid phases first separating from a molten magna are unlikely to be the most stable phases at ordinary temperatures and pressures.

If the rock deposits formed by cooling or other processes can undergo polymorphic transition to more stable states with the liberation of heat, and are on a sufficiently large scale to satisfy (1), from time to time such transformations may develop into weak detonation waves which will proceed with a definite stability of propagation to the limits of the deposit. It seems worth considering how far sudden transformations by weak detonative processes may account for the origin of certain earthquakes.

A NEW PHOTOGRAPHIC METHOD FOR STUDYING FAST TRANSIENT PHENOMENA

BY J. S. COURTNEY-PRATT

*Research Laboratory on the Physics and Chemistry of Rubbing Solids,
Department of Physical Chemistry, University of Cambridge*

This afternoon a number of methods have been discussed for following the course of propagation of explosions. Mechanical cameras have been described that have writing speeds of a few thousand metres per second. The speeds are limited by the ultimate strength and resistance to flexure of the moving parts, and even these speeds can only be attained at considerable sacrifice in overall light-gathering power.

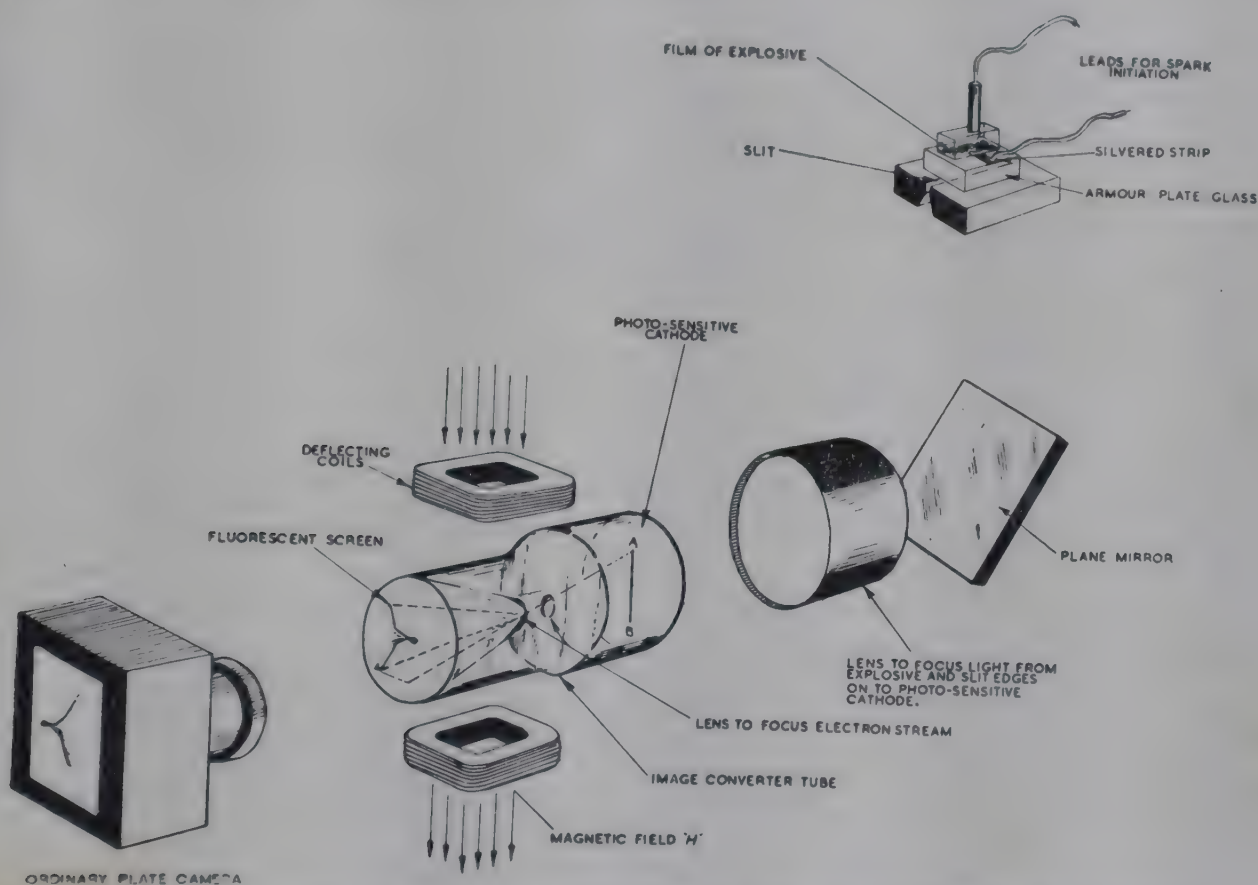


FIGURE 10. Schematic diagram of the apparatus used for photographing the birth and growth of explosive reactions.

A new method (Courtney-Pratt 1949) has been developed which makes use of the features of the image converter tube combined with a time-base deflexion system similar to any that could be used in a cathode-ray oscillograph. By avoiding entirely the use of mechanically moving parts the method is free of the inherent limitations of writing speed of all the earlier instruments. Vibration is eliminated, and this too, in general, results in better definition and resolution. Wide-aperture lenses can be used without difficulty, and the overall intensity compares favourably with that of the best mechanical cameras.

Suppose the photosensitive surface of the image converter tube, with its axis horizontal, is illuminated along a line AB in, say, the vertical plane as shown in figure 10. If the illumination is uniform, then with a standard tube there will be

a bright line CD of uniform intensity on the fluorescent screen. If an electric or magnetic field is introduced to act on the electron stream between the focusing electron lens system and the fluorescent screen, then the position of the line CD on the screen can be altered at will.

Suppose then that instead of uniform illumination of all points of AB , a bright point image is made to traverse linearly along AB , and further that there is a vertical magnetic field, or horizontal electric field, applied to the tube which also is changing linearly with time. The line then traced out on the fluorescent screen will be a diagonal, and the inclination of this diagonal will be a measure of the relative velocities of the image along AB and the rate of sweep of the time base.

Similarly, if the luminous image corresponding to any event is caused to move along AB while there is a changing deflecting field applied to the tube, there will be a line depicted on the fluorescent screen which is a function of the event and of the time-base sweep rate. This trace can be photographed with any ordinary stationary camera used with a time exposure equal to or greater than the duration of the event.

Figure 9*d*, plate 2, already described by Dr Bowden, shows the trace of detonating lead azide as the luminous explosion front travels a distance of a few millimetres away from the point of initiation at a velocity of 2000 m./sec.

The resolution on the fluorescent screen of the present tubes is better than 20 lines/mm. Figure 11, plate 3, is a photograph of the image of a page of print transmitted through the tube. It gives an indication of the resolution and freedom from serious distortion.

Sweep speeds greater than 10^5 m./sec. have been used so that the time resolution, i.e. the minimum interval of time between two events which may just be resolved is better than 10^{-9} sec.

Refer again to figure 9*d*. The light output accompanying the decomposition of any given grain of lead azide can be measured from the spread of the trace in the direction of sweep. Under the conditions of the experiments this duration is surprisingly short. The values for all the traces measured lay within the range 7×10^{-8} sec. $\pm 30\%$. The rise of light output is a feature which is distinguishable with an accuracy of about 10^{-8} sec. The brevity of this light pulse—or rather the sharpness of its leading edge—was used to demonstrate the time resolution of the method. Half the light from the decomposition of a given grain of lead azide was used to form a direct image on the photosensitive surface. The other half of the light was made to traverse a path 2.62 m. longer than the direct path and formed a second image on the photosensitive surface. The delay in arrival of the light forming the second image could be measured as a significant displacement of the first and second fluorescent images relative to one another (see figure 12). The resolution was here limited by the sharpness of the rising edge of the light pulse rather than by the characteristics of the image converter tube and associated gear, but was by this experiment shown to be certainly better than 10^{-8} sec.

The uses of the image converter tube in fast photography are not confined only to applications giving information with greater resolution but similar in form to that given by drum and mirror cameras. Single conventional photographs may be taken using an image converter tube and applying the supply voltage as a pulse of short

duration. Alternatively, a grid or control electrode can be incorporated and voltage pulses more conveniently applied to this electrode. The exposures can be as short at least as the best that are possible with flash photography and Kerr cells, and there is not the light loss unavoidably associated with the latter. A method has been developed for taking a series of exposures with a frame interval of the order of a microsecond. Separation of the individual frames is accomplished by applying suitable deflecting fields to the image converter tube. These methods may all be simply adapted for use with X-ray illumination.

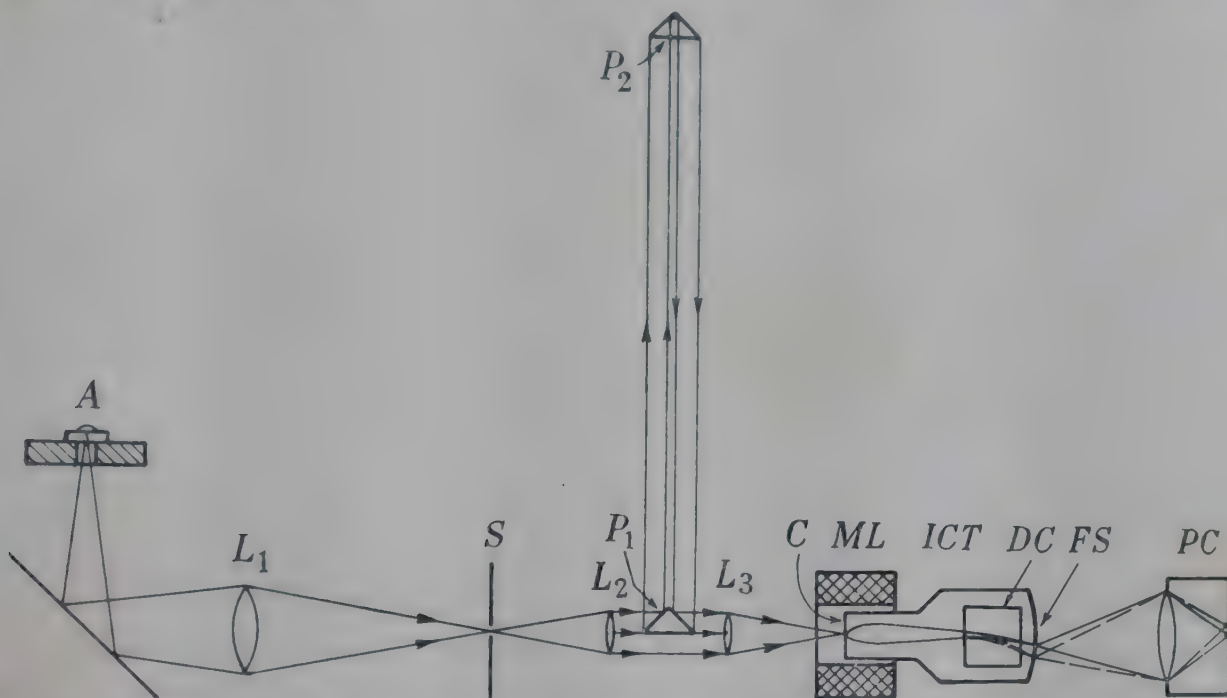


FIGURE 12. Diagram of the layout for testing the time resolution. The apparatus is capable of detecting the delay in arrival of a light pulse which has been caused to traverse a path that is 2.62 m. longer than the direct path. Light from the detonation of lead azide at *A* is focused on the slit *S*. Half the light passing through the slit goes directly through the lenses *L*₂ and *L*₃ to form the first image on the cathode *C* of the image converter tube *ICT*. The other half of the light is reflected from the face of prism *P*₁ to prism *P*₂ and back, and forms the second image on the cathode *C*. The electron streams focused by the magnetic lens *ML* and deflected by different amounts by the changing field from the coils *DC* produce fluorescent images on the screen *FS*. Both the fluorescent images are photographed by the plate camera *PC*.

The range and versatility of the methods make them applicable to the general study of fast movement, so that they can prove useful in the investigation of such problems as the detonation of explosives, combustion in engines, projectile and rocket flight, projectile target penetration, spark and discharge phenomena, and schlieren studies of shock waves.

SOME COMMENTS ON THE CONTRIBUTION BY DR COURTNEY-PRATT

By A. R. UBBELOHDE

The photographic observations on the initiation of detonation in lead azide agree with less direct methods which all lead to the same conclusion. Detonation in lead azide is normally initiated by a process which seems quite distinct from 'burning

to detonation'. But our experiments showed that when the air surrounding the crystals of lead azide was replaced by liquids of high boiling point, so as to change the Γ of the medium, it was possible to initiate lead azide by a 'burning' (self-heating) process. It would be very interesting to test this conclusion by the direct photographic technique.

THE REDUCTION IN REACTION ENERGY IN CASES OF INCOMPLETE CONFINEMENT

BY S. PATERSON

Research Department, I.C.I. Ltd., Nobel Division, Stevenston, Ayrshire

We regard the detonation wave as comprising

- (a) a non-reactive shock,
- (b) a reaction zone in which the explosive is transformed into products in chemical equilibrium,
- (c) a rarefaction in which these products expand to atmospheric pressure.

The Chapman-Jouguet layer forms the posterior surface of the *steady zone*, i.e. the region in which conditions remain unchanging from the standpoint of an observer accompanying the wave. In a cartridge under perfect lateral confinement, as Dr Penney has indicated, the steady zone and reaction zone coincide: in other words, the Chapman-Jouguet condition is to be applied at the point where equilibrium is reached. This can be sufficiently established. In the general case, where confinement is not perfect, we can still say that the steady zone falls entirely within the reaction zone, since a non-reactive rarefaction cannot be steady. But the converse does not necessarily follow; part of the reaction may take place outwith the steady zone. In this event, the velocity will tend to be reduced for two reasons: (a) the overall reduction in pressure, and (b) the fact that part of the chemical energy released is not available to support the wave. I think I am right in saying that Professor Jones's theory takes account of only the first factor; although the second may be of importance, perhaps of comparable importance in some cases. I believe Dr Devonshire has considered this question.

HIGH AND LOW DETONATION VELOCITY REGIMES IN CONDENSED EXPLOSIVES

BY J. TAYLOR

Research Department, I.C.I. Ltd., Nobel Division, Stevenston, Ayrshire

I would like to bring before your notice an extremely interesting and important point which has not so far been dealt with in the discussion, namely, the existence of high- and low-velocity regions in the detonation of condensed explosives. It has been known for many years that some liquid explosives, such as nitroglycerine and also gelatinous explosives like gelignites, are capable of detonating at one of two velocities depending on the strength of the initiation. Heavy priming gives velocities in a region round 7000 m./sec., and ordinary relatively weak priming

yields velocities in a region round 2000 m./sec. Intermediate velocities are not obtained, and there is a discontinuity below the high- and low-velocity regimes. The low velocity can be quite stable and propagate over long distances. In large-diameter cartridges, however, high-velocity propagation alone is possible. The explosives industry, with which I am connected, has been particularly interested in developing the hydrodynamic theory of detonation, so that quantitative calculations of the limiting velocity can be made for commercial explosives and the values calculated compared with practical results. In most cases the agreement has been surprisingly good, and the theory has provided an extremely useful means for the quantitative design of explosives. The high- and low-velocity regime, however, introduces difficulties urgently in need of resolution. It would appear that the method of decomposition of the explosive in the two cases is quite different, and we believe that the low-velocity region is explainable by deflagration on the surface of the grains or films of the explosive, initiated by hot spots produced by the heating of the interstitial gases, whereas the high-velocity region is produced by a bulk thermal decomposition throughout the explosive material induced by extremely high pressures.

There is still, however, a great deal to be done in this connexion, and I think the position is that the development of the theory of detonation has gone as far as it can using the 'mechanical' methods of mathematicians and physicists and now requires careful attention and research by the physical chemists to whom I would recommend this fascinating problem.

THE FADING OF DETONATION IN SOLID EXPLOSIVES

BY O. A. GURTON

Research Department, I.C.I. Ltd., Nobel Division, Stevenston, Ayrshire

In connexion with the experiments on the 'carry-over' of the impulse from the detonation of a priming charge into the secondary charge as described by Dr Evans, it should be pointed out that a glass plate placed in front of the charge may fracture at more than 5300 m./sec. and consequently obscure the detonation front (E. Jones 1928).

The absence of fading in an over-primed charge in the case described is not of course typical of detonating explosives. For example, a coarse-grained explosive at density 1.05 g./cm.³ may be set off at a velocity greater than its stable velocity by a number of priming charges, some of which detonate at a higher speed than the initial velocity of the explosive, and some which detonate at a lower speed. This has been shown by I. G. Cumming in some recent high-speed flame photographs. He has also been able to differentiate the curved traces and show how the velocity falls to a steady value. Below a certain diameter the fading continues, no steady velocity is reached, and the detonation fails to propagate. This complete fading of a detonation is illustrated by the high speed flame photograph shown in figure 13, plate 3. However (as is shown in figure 14) the rate of fading falls rapidly and the velocity tends to remain at the expected steady value. But the energy release is

insufficient, and the fading rate increases again leaving a point of inflexion in the velocity distance curve corresponding to the stable velocity to be expected. In some cases the whole process may be repeated during the continued fading, and another point of inflexion may be obtained at the expected low velocity of detonation. It is clearly impossible to ascribe this relatively complex behaviour of the detonation front in fading to effects outside the charge.

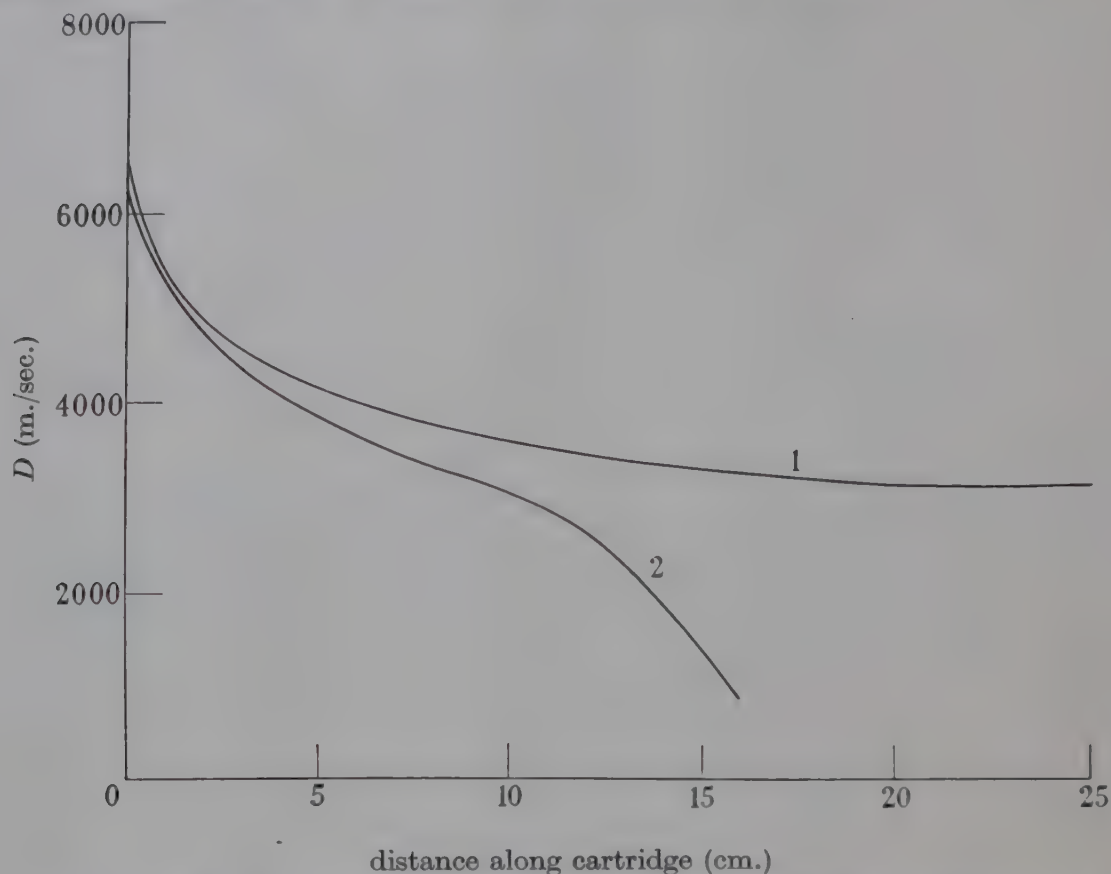


FIGURE 14. Build-down and fading of detonation in ballistite. $\Delta = 1.05 \text{ g./cm.}^3$.
Curve 1, $d = 2.54 \text{ cm.}$; curve 2, $d = 2.27 \text{ cm.}$

REFERENCES

- Bone, W. A. & Fraser, R. P. 1931 *Phil. Trans. A*, **230**, 363.
 Bowden, F. P., Mulcahy, M. F. R., Vines, R. G. & Yoffe, A. D. 1947 *Proc. Roy. Soc. A*, **188**, 311.
 Bowden, F. P. & Gurton, O. A. 1949 *Proc. Roy. Soc. A*, **198**, 337 and 350.
 Bowden, F. P., Stone, M. A. & Tudor, G. K. 1947 *Proc. Roy. Soc. A*, **188**, 329.
 Chapman, D. L. 1899 *Phil. Mag.* **47**, 90.
 Chree, C. 1889 *Proc. Camb. Phil. Soc.* **14**, 250.
 Copp, J. L. & Ubbelohde, A. R. 1948 *Trans. Faraday Soc.* **44**, 658.
 Courtney-Pratt, J. S. 1949 *Research*, **2**, 287.
 Cybulski, W. B., Payman, W. & Woodhead, D. W. 1949 *Proc. Roy. Soc. A*, **197**, 51.
 Davies, R. M. 1948 *Phil. Trans. A*, **240**, 375.
 Eyring, H., Powell, R. E., Duffey, G. H. & Parlin, R. B. 1949 *Chem. Rev.* **45**, 69.
 Gaydon, A. G. & Wolfhard, H. G. 1949 *Proc. Roy. Soc. A*, **196**, 105; **199**, 89.
 Geiling, S. & Richter, H. 1949 *Acta Crystallogr.* **2**, 305.
 Gray, P. & Yoffe, A. D. 1950 *Proc. Roy. Soc. A*, **200**, 14.
 Hicks, B. L. & Wasserman, R. H. 1947 *J. Appl. Phys.* **18**, 891.
 Hill, R. & Pack, D. C. 1947 *Proc. Roy. Soc. A*, **191**, 524.
 Hopkinson, B. 1914 *Phil. Trans. A*, **213**, 437.

- Jones, E. 1928 *Proc. Roy. Soc. A*, **120**, 603.
- Jones, H. 1947 *Proc. Roy. Soc. A*, **189**, 415.
- Jones, H. 1949 *Third symposium on combustion, flame and explosion phenomena*. Madison, Wisconsin, U.S.A. Baltimore: Williams & Wilkins.
- Jones, H. & Miller, A. R. 1948 *Proc. Roy. Soc. A*, **194**, 480.
- Jouguet, E. 1901 *C.R. Acad. Sci., Paris*, **132**, 673.
- Jouguet, [E.] 1905-6 *J. Math.* (6), **1**, 347.
- Kistakowsky, G. B. & Breit Wilson, E. 1941 *O.S.R.D. Report*, no. 114.
- Langweiler, H. 1938 *Z. tech. Phys.* **19**, 271.
- Lewis, B. & von Elbe, G. 1938 *Combustion, flames and explosives in gases*. Cambridge University Press.
- Love, A. E. H. 1934 *Mathematical theory of elasticity*. Cambridge University Press.
- Marshall, A. 1932 *Explosives*, 2nd ed. London: J. and A. Churchill.
- Owen, J. D. & Davies, R. M. 1949 *Nature*, **164**, 752.
- Paterson, S. 1950 *Research*, **2**, 99.
- Pochhammer, L. 1876 *J. reine angew. Math.* **81**, 324.
- Prandtl, L. 1907 *Phys. Z.* **8**, 23.
- Road Research Laboratory 1944 Ministry of Supply unpublished report.
- Robertson, A. J. B. & Yoffe, A. D. 1948 *Nature*, **161**, 806.
- Taylor, G. I. 1941 'Analysis of explosion of a long cylindrical bomb detonated at one end.' Ministry of Home Security unpublished report.
- Taylor, G. I. 1946 *Proc. Roy. Soc. A*, **186**, 273.
- Taylor, Sir Geoffrey 1950a *Proc. Roy. Soc. A*, **200**, 235.
- Taylor, Sir Geoffrey 1950b *Proc. Roy. Soc. A*, **201**, 159.
- Taylor, G. I. & Jones, H. 1942 Ministry of Supply unpublished report.
- Taylor, G. I. & Maccoll, J. W. 1933 *Proc. Roy. Soc. A*, **139**, 278.
- Ubbelohde, A. R. 1937 *Trans. Faraday Soc.* **33**, 1198.
- Ubbelohde, A. R. 1949 *Third symposium on combustion, flame and explosion phenomena*, p. 566. Madison, Wisconsin, U.S.A. Baltimore: Williams & Wilkins.
- von Neumann, J. 1942 *O.S.R.D. Report*, no. 1140.

A study of sensitized explosions

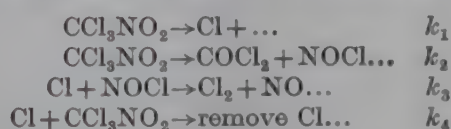
X. The kinetics of decomposition of chloropicrin and of the hydrogen-oxygen and hydrogen-chlorine reactions sensitized by chloropicrin

By P. G. ASHMORE AND R. G. W. NORRISH, F.R.S.

Laboratory of Physical Chemistry, University of Cambridge

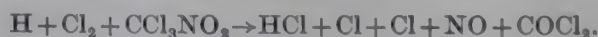
(Received 9 May 1950)

A kinetic scheme is suggested for the decomposition of chloropicrin which is shown to be in agreement with the results of the work of Smith & Steacie (1938*a, b*), and which provides an explanation of the sensitizing action of chloropicrin on the hydrogen-oxygen and hydrogen-chlorine reactions described in parts VIII and IX (Ashmore & Norrish 1950*a, b*):



Estimates of the relative concentrations of chlorine atoms at various stages of the decomposition at different temperatures are made. The concentration of chlorine atoms passes through a maximum, and at temperatures above 340° C this maximum is probably reached within 1 msec. It occurs earlier, and is greater in magnitude, the higher the chloropicrin pressure at a given temperature or the higher the temperature for a given concentration of chloropicrin. These variations are shown to provide a direct explanation of the results for the range 340 to 400° C described in part IX, and to enable the results of part VIII to be explained by a modification of the scheme proposed by Dainton & Norrish (1941) for the nitrosyl chloride-hydrogen-oxygen reaction. This modification consists of an allowance for the variation in the initial number of chain-centres arriving in the reaction vessel, the variation being due to changes in the precise location within the entry tube of the maximum concentration of the chlorine atoms which initiate the chains.

Kinetic schemes are suggested for the hydrogen-chlorine-chloropicrin reaction in the temperature region 100 to 200° C. A straight chain scheme, with chains initiated from reaction k_1 , is combined with a thermal condition for ignition to give a satisfactory account of the variation of the ignition limits with concentration of chloropicrin, with proportion of reactants, and with temperature, and also to account for the course of the slow reaction below the ignition limit. A branched chain mechanism is also discussed and is shown to account for the results only if the branching reaction is of the form



The experimental results do not allow a decision to be made between these possible schemes.

INTRODUCTION

The experimental work of the previous two parts (Ashmore & Norrish 1950*a, b*) has established that when chloropicrin decomposes, an atom or radical is formed capable of initiating hydrogen-oxygen or hydrogen-chlorine chains. The chains thus started in sensitized mixtures are considerably modified by the nitrosyl chloride produced from the chloropicrin. The atom or radical also reacts rapidly with nitrosyl chloride to give relatively non-reactive end products.

These conclusions can be related to those drawn by Smith & Steacie (1938*a, b*) from their experimental work on the decomposition of chloropicrin. These investigators

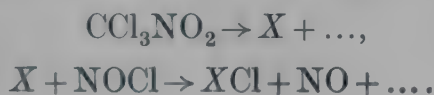
found that: (1) the decomposition of chloropicrin is a first order reaction, with a large energy of activation. The velocity constant is given by

$$k = 4.9 \times 10^{15} \times e^{-37,670/RT} \text{ sec.}^{-1};$$

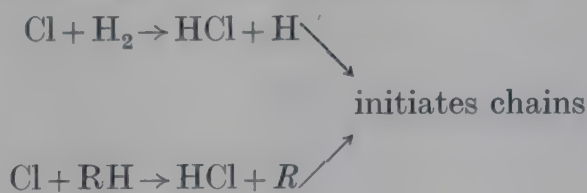
(2) the main products of decomposition are carbonyl chloride and nitrosyl chloride, with small amounts of nitric oxide and chlorine; (3) the nitric oxide and chlorine are present throughout the reaction in amounts which are not very different from those expected from the equilibrium $\text{NOCl} \rightleftharpoons \text{NO} + \frac{1}{2}\text{Cl}_2$. Smith & Steacie proved that this equilibrium could not be established from either side by the normal reactions, in the time available.

The agreement between the amount of decomposition predicted from the above value of k and the behaviour of sensitized mixtures above 300°C has already been indicated (part VIII, p. 469 and part IX, p. 483).

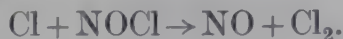
The postulated occurrence of an atom or radical X during the decomposition of chloropicrin suggests that the following reactions might occur in parallel with the main decomposition into carbonyl chloride and nitrosyl chloride



From an examination of the formula of chloropicrin, there are a number of possible structures for X . It might, for example, be Cl or COCl or CCl_3 . With nitrosyl chloride, Cl would yield NO and Cl_2 , COCl would yield CO , NO , and Cl_2 , while CCl_3 would probably give CCl_4 and NO (Dainton 1947). The fact that Smith & Steacie found an almost quantitative yield of carbonyl chloride, rather less than a quantitative yield of nitrosyl chloride, with the material balance substantially completed by nitric oxide and chlorine (with negligible amounts of other gases) is strongly in favour of X being a chlorine atom. Chlorine atoms can initiate hydrogen-oxygen, hydrogen-chlorine or hydrocarbon-oxygen chains by the reactions



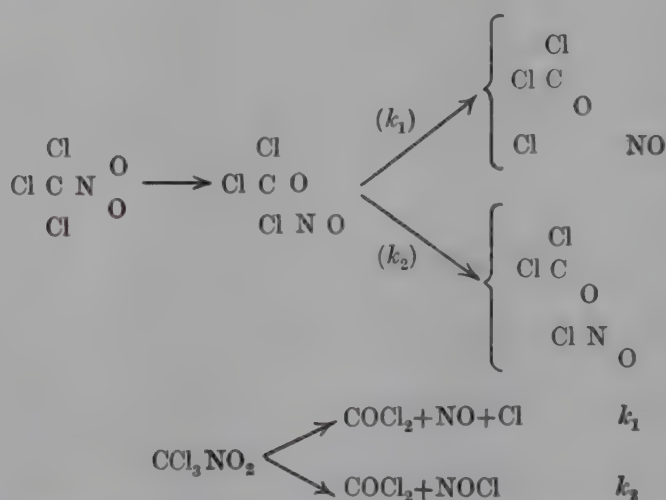
and would be removed by the reaction



It also seems reasonable to expect the formation of chlorine atoms by a mechanism similar to the second scheme proposed by Smith & Steacie. The molecules of carbonyl chloride and nitrosyl chloride which emerge from each elementary decomposition step will possess energy considerably in excess of the equipartition value for the temperature concerned. The decomposition requires a large energy of activation (~ 38 kcal.) and is exothermic to the extent of 17 kcal. as estimated from bond energies (Skinner 1945). A higher proportion of the product nitrosyl chloride molecules will possess energy in excess of the 38.5 kcal. necessary for dissociation into nitric oxide and a chlorine atom, then in normal nitrosyl chloride at that

temperature, so that the concentration of chlorine atoms may temporarily reach much higher values than in a system which contains the same number of 'normal' nitrosyl chloride molecules as have been formed from the chloropicrin.

It is impossible to tell at present whether the chlorine atom is formed directly from the activated complex formed by the chloropicrin, or from those nitrosyl chloride molecules (coming from the complex) which are sufficiently excited. Since most of the nitrosyl chloride molecules formed will not be sufficiently excited to dissociate, it seems better formally to retain two parallel modes of decomposition, remembering that they may pass through the same transition complex. The formation of this transition complex can be visualized as the simultaneous stretching and bending of the C—N bond. The formation process is favoured by the symmetry of the chloropicrin molecule.



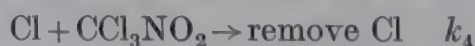
After the two parallel reactions k_1 and k_2 , there will follow a rapid interaction between the chlorine atom and the nitrosyl chloride formed in k_1 and k_2 respectively (Dainton, 1947)

$$\text{Cl} + \text{NOCl} \rightarrow \text{Cl}_2 + \text{NO} \quad k_3$$

• The reverse reaction, which has an activation energy of at least 19.5 kcal., will not be important at lower temperatures, but may be a source of chlorine atoms in the presence of excess chlorine at higher temperatures.

The reaction k_1 is slower than the main reaction k_2 because NOCl and not NO is in excess in the products. For every Cl atom formed, two molecules of nitric oxide and $\frac{k_2}{k_1} - 1$ of nitrosyl chloride result. This enables a rough estimate of the ratio k_1/k_2 to be obtained from the fact that Smith & Steacie found, during the course of decomposition, that the ratio of nitric oxide to nitrosyl chloride was about 15:85. Then $\frac{2k_1}{k_2 - k_1} = \frac{15}{85}$ or $\frac{k_1}{k_2} \sim \frac{1}{12}$. This is probably an upper estimate, because some of the nitric oxide would have been formed by dissociation of the nitrosyl chloride.

In view of reactions k_1 and k_2 , it appears that in addition to the reaction k_3 there may be another reaction which removes chlorine atoms

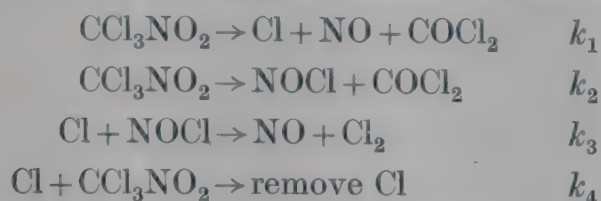


However, the experiments on the sensitized hydrogen-chlorine reaction described in part IX suggest that k_4 is very much less than k_3 at temperatures below 400°C.

In particular, the very different effects of chloropicrin and of nitrosyl chloride on the ignition limits, the 'backfire' phenomenon, and the strong inhibiting effect of nitrosyl chloride on the chloropicrin-sensitized ignition support this view.

The decomposition can therefore be considered as a main reaction (k_2) with a lesser side reaction (k_1) and subsequent interactions, one (k_3) strong and one (k_4) weak.

THE KINETICS OF THE DECOMPOSITION OF CHLOROPICRIN



where k_2 is large compared with k_1 or with k_4 $[\text{Cl}]$ and k_3 is large compared with k_4 . These reactions lead to a rate equation for $[\text{Cl}]$ which can be simplified to the following as soon as $k_3[\text{NOCl}]$ is greater than $k_4[\text{CCl}_3\text{NO}_2]$. (The subscript zero refers to initial concentration).

$$\frac{d[\text{Cl}]}{dt} = k_1[\text{CCl}_3\text{NO}_2]_0 e^{-k_2 t} - k_3[\text{Cl}][\text{CCl}_3\text{NO}_2]_0 \{1 - e^{-k_2 t}\}.$$

Let $e^{-k_2 t} = u$, and write $\frac{k_1}{k_2} [\text{CCl}_3\text{NO}_2]_0 = b$, $\frac{k_3}{k_2} [\text{CCl}_3\text{NO}_2]_0 = a$.

Substituting and integrating

$$\frac{[\text{Cl}]}{b} = \frac{\int_{u_1}^1 u^{-a} e^{au} du}{u_1^{-a} e^{au_1}} \text{ for } 0 < u_1 < 1.$$

The integration cannot be performed directly. The integral can be evaluated for any particular value of a by calculating $a^{-a} e^{au}$ from tables for a range of values of u , and applying Simpson's rule or performing a graphical summation.

The evaluation is difficult, however, for values of a exceeding 100. Fortunately, in the temperature range 300° to 400° C, studied in the hydrogen-oxygen-chloropicrin reaction, the values of a are sufficiently low to permit numerical calculations. This temperature range is also of interest in the hydrogen-chlorine-chloropicrin reaction (cf. figure 4, part IX), and in the next section an estimate is made of the comparative values of $[\text{Cl}]$ during the decomposition of chloropicrin between 300° and 400° C and the results are applied to the sensitized ignitions in this range. In the following section (p. 42) a somewhat different approach is made to the kinetics of the hydrogen-chlorine-chloropicrin reaction between 100° and 200° C, based upon the reaction scheme given above.

ESTIMATION OF THE COMPARATIVE VALUES OF $[\text{Cl}]$ DURING THE DECOMPOSITION OF CHLOROPICRIN BETWEEN 300° AND 400° C.

The parameter a is given by $a = \frac{k_3}{k_2} [\text{CCl}_3\text{NO}_2]_0$. The value of k_2 at any temperature can be taken as equal to k , the velocity constant for the decomposition of chloropicrin, and so determined from the formula given by Smith & Steacie (1938)

$k = 4.9 \times 10^{15} \times e^{-37670/RT}$ sec.⁻¹. The value of k_3 at any temperature can be estimated from a formula derived by Burns & Dainton (1950) from their work on the effect of nitrosyl chloride on the photochemical reaction between carbon monoxide and chlorine. This formula is $k_3 = 1.14 \times 10^{10} \times e^{-1060/RT}$ l./g.mole/sec. The concentrations of chloropicrin which are of interest range from 0.05 mm. to 3 mm. at 400° C (e.g. part VIII, figure 3 and part IX, figures 3 and 4). The values, at different temperatures, of k_2 , k_3 and a for $[\text{CCl}_3\text{NO}_2]_0 = 0.1$ mm., are shown in table 1.

TABLE 1

temperature (° C)	300	320	340	360	380	400
k_2 sec. ⁻¹	20.5	61.7	178	468	1176	2760
k_3 l./g.mole/sec.	4.50×10^9	4.65×10^9	4.79×10^9	4.93×10^9	5.05×10^9	5.16×10^9
a for 0.1 mm. CCl_3NO_2 (2.4×10^{-6} g.mole/l.)	527	181	65	25	10	4.5

For any other initial concentration of chloropicrin, a is by definition proportional to $[\text{CCl}_3\text{NO}_2]_0$.

These calculations of a refer to pressures of chloropicrin lower than those used to determine the value of k . But it has been shown that in the presence of foreign gases, particularly hydrogen, the partial pressure of reactant at which a first order velocity constant falls off is often much lower than in the absence of foreign gas. The calculations also refer to much higher temperatures than those at which k_2 was determined, but some evidence is given on p. 35 which shows that the values of k are of the correct order of magnitude.

The parameter b is given by the equation $b = \frac{k_1}{k_2} [\text{CCl}_3\text{NO}_2]_0$. b cannot be determined absolutely because k_1 is unknown. The ratio $\frac{k_1}{k_2}$ depends upon the way in which the potential energy of the activated complex $\text{CCl}_3\text{NO}_2^*$ is distributed among the product molecules NOCl and COCl_2 . The ratio has been estimated as $\sim \frac{1}{12}$. It will be assumed that this ratio is independent of temperature, and for convenience of calculation arbitrary values of b will be taken such that $b = 10$ for $[\text{CCl}_3\text{NO}_2]_0 = 0.1$ mm. at all temperatures, and is proportional to $[\text{CCl}_3\text{NO}_2]_0$ at all temperatures.

The quantity $u^{-a} e^{au}$ has been calculated for intervals of u of 0.01 between 0.1 and 1.0, with a equal to 1, 3, 5, 10, 20, 50, 100, and the integral has been evaluated by applying Simpson's rule. Corresponding values of $\frac{[\text{Cl}]}{b}$ have been evaluated.

From these calculations, and the relation $t = \frac{1}{k_2} \ln \frac{1}{u}$, estimates have been made of the comparative values of $[\text{Cl}]$ at any time after the decomposition of a given concentration of chloropicrin has started. The values of a and b are chosen appropriate to the temperature and concentration of chloropicrin.

Figure 1 shows how $[\text{Cl}]$ varies with time for various initial concentrations of chloropicrin, at a fixed temperature. Figure 2 shows how $[\text{Cl}]$ varies with time for various temperatures, with a constant initial concentration of chloropicrin.

From these figures the following conclusions are evident—(a) that $[Cl]$ passes through a maximum as the decomposition proceeds; (b) at a given temperature (figure 1) the maximum is higher and occurs earlier, the greater the initial concentration of chloropicrin. After a time, however, the concentration of chlorine atoms is lower for higher initial chloropicrin concentrations; (c) at a fixed initial concentration of chloropicrin (figure 2) the maximum is higher and occurs earlier, the higher the temperature. At later times, however, the concentration of chlorine atoms is lower, the greater the temperature.

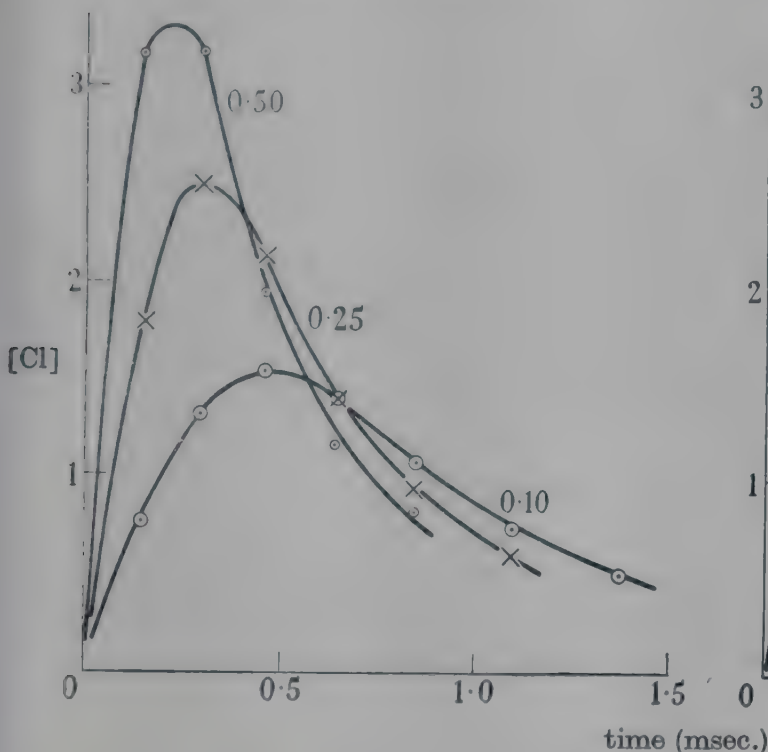


FIGURE 1. The variation of $[Cl]$ with time, for various pressures (mm.) of chloropicrin, at $365^{\circ}C$. $[Cl]$ in arbitrary units.

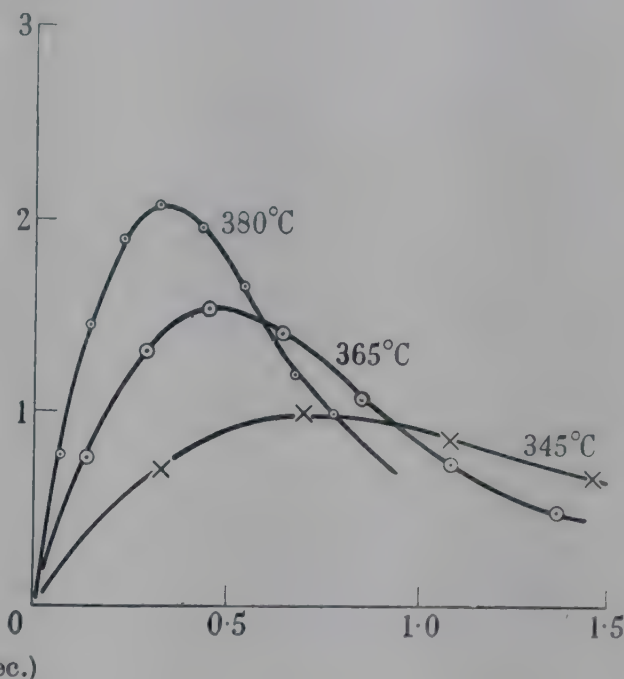


FIGURE 2. The variations of $[Cl]$ with time, for various temperatures, in the presence of 0.10 mm. of chloropicrin. $[Cl]$ in arbitrary units.

APPLICATION OF THE RESULTS SHOWN IN FIGURES 1 AND 2 TO THE SENSITIZED IGNITIONS, 300 TO $400^{\circ}C$.

(a) The hydrogen-oxygen ignitions

The general discussion of the experimental results of part VIII has established that the chloropicrin-sensitized reaction between hydrogen and oxygen behaves as though the nitrosyl chloride formed by the decomposition of the chloropicrin controls the branching and so determines the ignition limits. The induction periods which would be expected if the nitrosyl chloride alone determined the course of the reactions leading to ignition are, however, considerably shortened, by the appearance of free chlorine atoms during the rapid decomposition of the chloropicrin. In the presence of hydrogen, these chlorine atoms will react to give hydrogen atoms which will start the hydrogen-oxygen chains



It has been estimated in part VIII that the gases take several milliseconds to flow down the heated entry tube of the vessel, so that the curves of figures 1 and 2 show that the maximum value of $[Cl]$ occurs in the entry tube. The rate of production of chain centres is greatest at the maximum value of $[Cl]$, and in the narrow entry tubes where the branching introduced by the nitrosyl chloride is outweighed by the chain terminating processes (Dainton & Norrish 1941) the concentration of centres n must follow curves very similar to those of figure 2 during the flow of gases into the vessel. The concentration of chain-centres arriving in the vessel will depend upon the maximum concentration of hydrogen atoms formed, the (negative) value of the net branching factor ϕ in the narrow (6 mm.) entry tube, and the length of time spent in the tube. In the wider entry tubes used with quartz vessels (10 mm. diameter) ϕ may be positive, so that the chains can multiply from the earliest stages within the entry tube and give very short induction periods with mixtures which in Pyrex vessels give induction periods of several seconds (see part VIII, p. 470).

In their theoretical treatments of sensitized ignitions, most investigators assume that the concentration of chain centres present in the reaction vessel immediately after admission of the gases is so low that it can be neglected. The ensuing concentration of centres n is assumed to conform to the growth equation

$$\frac{dn}{dt} = \theta + \phi n,$$

where θ is the rate of initiation of centres and ϕ is the net branching factor (in the present treatment, quadratic branching and termination will not be included). The induction periods end when the concentration of chain centres reaches a critical value n_c , and it has been shown that if $n = 0$ when $t = 0$,

$$\tau = \frac{1}{\phi} \ln \left(1 + \frac{n_c \phi}{\theta} \right).$$

If the initial concentration of centres is assumed to be n_1 instead of zero, however, the induction period becomes

$$\tau = \frac{1}{\phi} \ln \left(\frac{\theta + n_c \phi}{\theta + n_1 \phi} \right).$$

The change in the induction period as the initial concentration of centres varies can be seen from a consideration of figure 3*a*. In this figure, n is plotted against t for particular values of θ and ϕ , the scales being arbitrary. When the initial concentration of centres is zero, the length of the induction period is represented by $A_0 B_0$. When the initial concentration is n_1 the induction period is $A_1 B_1$ and so on.

The figure can be used to illustrate the differences in induction periods when mixtures identical in reactant composition are sensitized first by chloropierin and then by the same amount of nitrosyl chloride. It has been shown that the mixture which reaches the reaction vessel behaves in both cases as if the sensitizer were nitrosyl chloride, and because one molecule of chloropierin yields one of nitrosyl chloride (less the small amount decomposed to nitric oxide and chlorine) the values of θ and ϕ in the vessel are the same for the chloropierin-sensitized and the nitrosyl chloride-sensitized mixtures. The rate of growth of n in the vessel can therefore be

represented by the one curve of figure 3a for both sensitized mixtures. The induction period of the nitrosyl chloride-sensitized mixtures, in which the initial concentration of centres is assumed to be zero, can therefore be represented by A_0B_0 , and that of the chloropicrin-sensitized mixture by A_1B_1 . A similar difference in induction periods will occur for each pair of sensitized mixtures, so that in a figure showing the variation of induction periods with sensitizer concentration, the curve for chloropicrin will always lie below that for nitrosyl chloride. It may be concluded that this is the explanation of the relative positions of the bottom and top curves of figure 12 of part VIII.

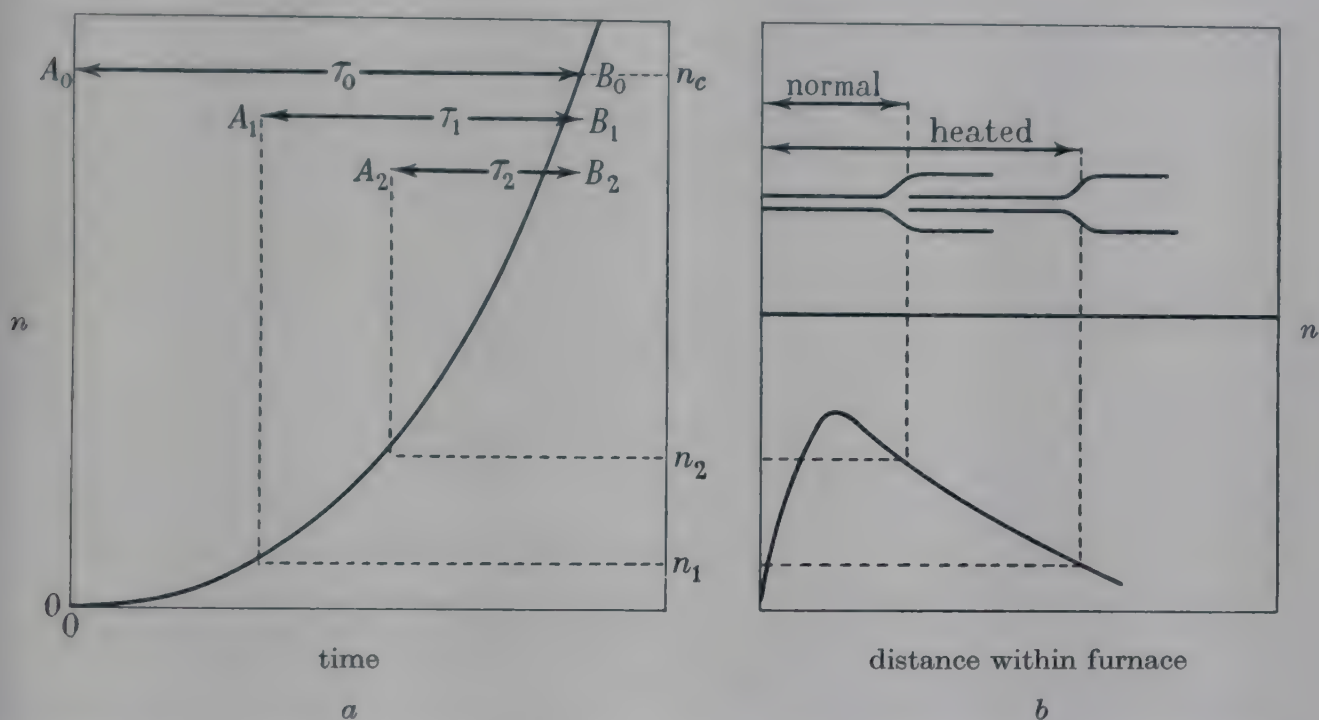


FIGURE 3a. Illustration of the effect of varying the initial concentration of centres upon the induction period τ of any sensitized mixture. *b* The effect of varying the heated length of the entry tube of vessel *B* upon the induction period of chloropicrin sensitized mixtures.

In a similar manner, the concentration of centres n arriving in the vessel will be lower after passing through a longer length of heated tube. Hence, it may be seen from figures 3b and 3a that the induction period τ_1 for the 'heated' tube will be longer than that for the 'normal' tube of vessel *B* as found experimentally.

(b) The hydrogen-chlorine-chloropicrin ignitions

In this reaction, the hydrogen atoms formed in the propagating step $\text{Cl} + \text{H}_2 \rightarrow \text{HCl} + \text{H}$ will be replaced very rapidly by chlorine atoms, by the fast reaction $\text{H} + \text{Cl}_2 \rightarrow \text{HCl} + \text{Cl}$. It is therefore reasonable to apply the curves shown in figures 1 and 2 directly to this reaction. The variations in $[\text{Cl}]$ shown in figure 2 are in complete agreement with the experimental results found for the hydrogen-chlorine-chloropicrin reaction in the temperature range of 300° to 400° C. The gases take a few milliseconds to flow into the vessel; during this period, $[\text{Cl}]$ passes through a maximum, and, by the time the gases reach the vessel, $[\text{Cl}]$ falls to a value which is lower the greater the temperature. The rate of reaction with a given concentration of reactants is proportional to $[\text{Cl}]$ and is therefore lower, in this temperature range,

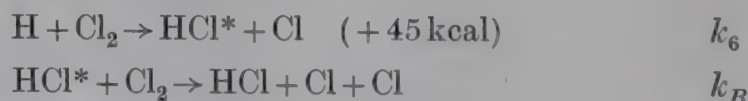
the greater the temperature, so that the concentration of reactants must be increased at higher temperatures in order to reach the critical rate required for ignition. This means that the minimum pressure for ignition rises with increasing temperature, and explains the occurrence of the rising boundaries shown in figure 4 of part IX. Moreover, the shorter the heated entry tube, the nearer would $[Cl]$ be to the maximum value for each particular temperature, and hence the minimum pressure for ignition would be lower. This is also observed.

The effect of increasing the chloropicrin concentration at any temperature between 300° and $400^\circ C$ is to decrease the value of $[Cl]$ at the time the gases reach the vessel (figure 1). Following the same argument as above, it would be expected that the ignition boundary at, say, $350^\circ C$ would be higher the higher the chloropicrin pressure. This is also observed (figure 4, part IX). In view of the complexity of the reactions involved, it does not seem profitable to discuss further the hydrogen-chlorine-chloropicrin systems in this temperature range (300° to $400^\circ C$).

THE KINETICS OF THE HYDROGEN-CHLORINE-CHLOROPICRIN REACTION IN THE TEMPERATURE RANGE 100° TO $200^\circ C$.

The reactions normally postulated for the thermal hydrogen-chlorine reaction chain probably occur in the reaction sensitized by chloropicrin, in the temperature region of 100° to $200^\circ C$. The marked effect of chloropicrin may then be attributed to its ability to initiate many more chains at these low temperatures than would normally be started in a given time, provided that it does not introduce any too effective chain-terminating steps which would materially shorten the chains.

It is possible that the chloropicrin molecules introduce some step into the hydrogen-chlorine chains which is a branching reaction. Semenoff (1935) suggested that a form of branching could occur in the unsensitized reaction between hydrogen and chlorine by the consecutive steps



If k_B is replaced by the reaction k_C

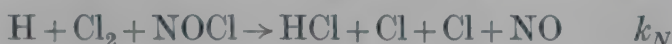


the sensitization could be ascribed to a 'chain-stimulated initiation' which is formally a branching step, because the concentration of HCl^* depends on the concentration of chain centres. Reaction k_C would occur at lower temperatures than reaction k_B because the activation energy for the decomposition of chloropicrin is 38 kcal. compared with 57 kcal. for chlorine. There are some features of the sensitized ignitions occurring between 100° and $200^\circ C$ which point to a branched chain mechanism, for example the abrupt transition from slow reaction to ignition (part IX, figure 7) and the occurrence of an induction period at low pressures of sensitizer (part IX, figure 5). However, these features do not exclude a straight-chain thermal ignition (Dainton 1942). Moreover, strict application of the branched-chain theories of ignitions (whether thermal or isothermal) meets with the difficulty that the net branching

factor ϕ decreases with time, due to the increase in the rate of chain termination by the nitrosyl chloride which is continuously formed by reaction k_2 . The chief interest lies, however, in the initial stages of the reaction, when the concentration of nitrosyl chloride is low. An expression for the initial values of ϕ can be derived (see appendix B) and combined with the isothermal condition for ignition $\phi = 0$ to describe the ignition limits in a satisfactory manner, provided that reaction k_6 followed by k_C is replaced by a single branching step $k_{C'}$ involving a three-body collision

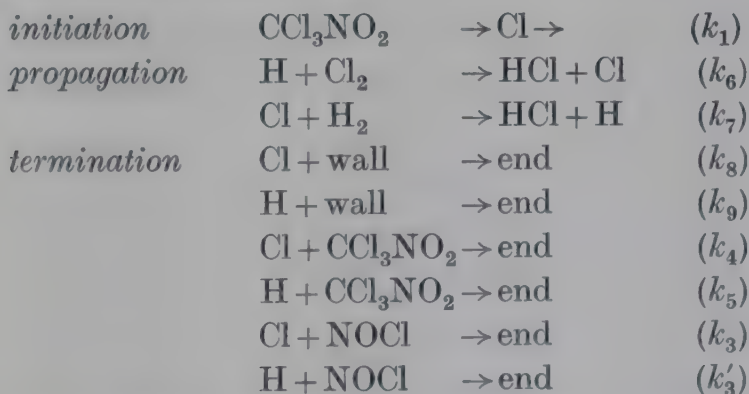


It may also be mentioned that the results obtained with mixtures of hydrogen, chlorine, and nitrosyl chloride (part IX, figure 11) can only be explained on an isothermal branched-chain condition for ignition by introducing a similar three-body collision



A straight-chain kinetic scheme is outlined below and is combined with a thermal condition of ignition to give equations which describe the experimental results as adequately as do those derived from the branched chain scheme. Centres are formed by reaction k_1 , and the rate of reaction increases rapidly as the concentration of centres builds up to a maximum. When sufficient nitrosyl chloride is formed to make reaction k_3 important, the concentration of centres and the rate of reaction will fall. The concentration of centres will follow curves similar to those shown in figure 1, except that between 100 and 200°C the time scale will be in seconds rather than milliseconds. The maximum rate of reaction will be greater and will be reached in a shorter time, the higher the chloropicrin concentration.

The straight-chain reaction scheme applicable to the reaction at low temperature is



It is known that $k_6 > k_7$ (Morris & Pease 1935) and $k_3 \gg k_4$. The experimental results shown in figure 11 of part IX suggest that at low temperatures $k_3[\text{NOCl}] \gg k_8$ as soon as $[\text{NOCl}]$ is ~ 0.1 mm. k'_3 would be expected to be of the same magnitude as k_3 , and k_5 as k_4 . The above scheme gives an equation for the maximum rate while $[\text{NOCl}]$ is very small which can be simplified, using the above assumptions (see appendix A), to

$$\frac{d[\text{HCl}]}{dt} = \frac{2k_1 k_6 k_7 [\text{Cl}_2] [\text{H}_2] [\text{CCl}_3\text{NO}_2]}{k_7 [\text{H}_2] \{k_9 + k_5 [\text{CCl}_3\text{NO}_2]\}} + k_6 [\text{Cl}_2] \{k_8 + k_4 [\text{CCl}_3\text{NO}_2]\}. \quad (1)$$

The condition for ignition is that this maximum rate should exceed a critical value which depends upon the ability of the system to conduct away the heat liberated in

the reaction and to maintain isothermal conditions. It has been shown (Frank-Kamenetzki 1939) that the condition is

$$r^2 \frac{Q}{\lambda} \frac{E}{RT^2} \text{rate} = \text{constant}, \quad (2)$$

where the rate of reaction can be written in the form

$$\text{rate} = k e^{-E/RT} (p_1^{n_1} p_2^{n_2} \dots). \quad (3)$$

Q is the heat of reaction, E is the energy of activation, λ is the thermal conductivity of the system, r is the radius of the vessel, and the constant in (2) takes the values 2.00 for a cylindrical and 3.32 for a spherical vessel. Q , λ , r are constant for a given proportion of reactions and a given vessel. p_1 is the pressure of reactant 1, and n_1 is a constant representing the order with respect to reactant 1, and so on. Application of the condition (2) to equation (1) then leads to the following results:

(a) *Variation of ignition limit with chloropicrin concentration (T constant, and proportion of reactants constant)*

The rate can be written
$$\frac{d[\text{HCl}]}{dt} = \frac{[\text{CCl}_3\text{NO}_2] p}{\frac{A}{p} + B[\text{CCl}_3\text{NO}_2]}, \quad (4)$$

where A and B are now constants. The condition for ignition is that this rate should exceed some critical value C , and hence, for ignition

$$[\text{CCl}_3\text{NO}_2] \{p - BC\} \geq \frac{AC}{p}. \quad (5)$$

The variation of p with $[\text{CCl}_3\text{NO}_2]$ will clearly, as a first approximation, be a hyperbola with axes $[\text{CCl}_3\text{NO}_2] = 0$ and $p = BC$. This equation is tested by plotting p against $\frac{1000}{[\text{CCl}_3\text{NO}_2] p}$ (figure 4) and it is seen that the points so plotted are a reasonable fit to a straight line.

(b) *Variation of ignition limit with the proportion of reactants ($[\text{CCl}_3\text{NO}_2]$ and T constant)*

Let $[\text{H}_2] = yp$ and $[\text{Cl}_2] = (1-y)p$.

The rate can be written
$$\frac{d[\text{HCl}]}{dt} = \frac{y(1-y)p}{Dy + E}, \quad (6)$$

where E is constant and D is a function of p^{-1} under these conditions.

The critical condition for ignition is then

$$\frac{y(1-y)p}{Dy + E} \geq C, \quad (7)$$

where C is itself a function of y because it depends *inter alia* on the thermal conductivity of the system. The variation of the thermal conductivity of the mixture is such that $\frac{dC}{dy}$ is almost certainly positive. (*I.C.T.* 5, p. 214). It can be shown from

this that p passes through a minimum value at $y = \frac{1}{2} \left(1 - \frac{\theta}{p} \right)$ where θ is positive. This is in agreement with the experimental results in part IX (figure 6), where the minimum value of p occurs at $y = 0.30$, i.e. $\frac{\theta}{p} = +0.40$, but θ is too complex to make any further deductions.

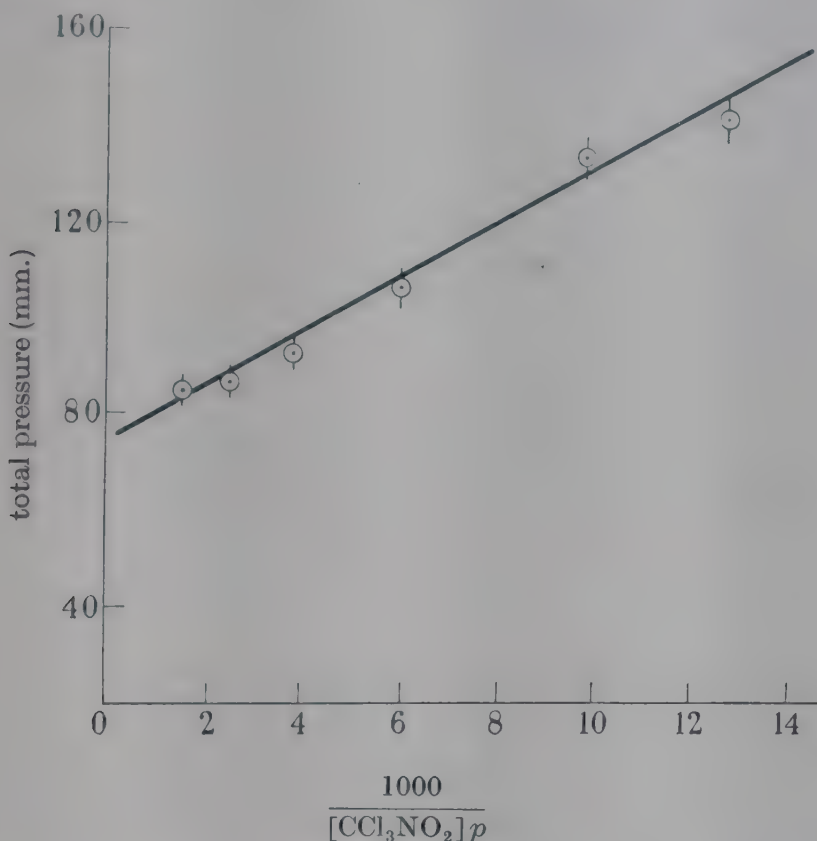


FIGURE 4. Plot of total pressure $p(\text{H}_2 + \text{Cl}_2)$ against the function $\frac{1000}{[\text{CCl}_3\text{NO}_2] p}$ for the vessel B at 162°C . p is the minimum ignition pressure (mm.)

(c) Variation of ignition limit with temperature ($[\text{CCl}_3\text{NO}_2]$ and proportion of reactants constant).

The rate equation (3) can be written, for stoichiometric mixtures when $y = \frac{1}{2}$ as

$$\frac{d[\text{HCl}]}{dt} = \frac{k_1 k_6 k_7 [\text{CCl}_3\text{NO}_2] p}{k_7 k_9 + k_6 k_8 + \{k_5 k_7 + k_4 k_6\} [\text{CCl}_3\text{NO}_2]}. \quad (8)$$

If the second term in the denominator of (8) predominates at high chloropierin concentrations

$$\frac{d[\text{HCl}]}{dt} = \frac{k_7 k_6 k_1 p}{k_7 k_5 + k_6 k_4} \quad (9)$$

E_6 and E_7 are not large and E_4 and E_5 are probably not very different. If it is further assumed that $E_7 + E_5 \sim E_6 + E_4$ and $k_7 k_5 \sim k_6 k_4$, a very rough equation for the rate is

$$\frac{d[\text{HCl}]}{dt} = \frac{k_7 k_1 p}{2k_4} = kp \exp \left[-\frac{E_1 + E_7 - E_4}{RT} \right] \quad (10)$$

giving as the condition for ignition

$$\log_{10} \frac{p}{T^2} = \frac{E_1 + E_7 - E_4}{2.303RT} + \text{constant.} \quad (11)$$

The plots of $\log_{10} \frac{p}{T^2}$ against $\frac{1}{T}$ for the single entry vessel (with chloropicrin equal to 2, 10, and 18 mm.) and for the double entry vessel (with chloropicrin equal to 0.55 mm.) are shown in figure 5.

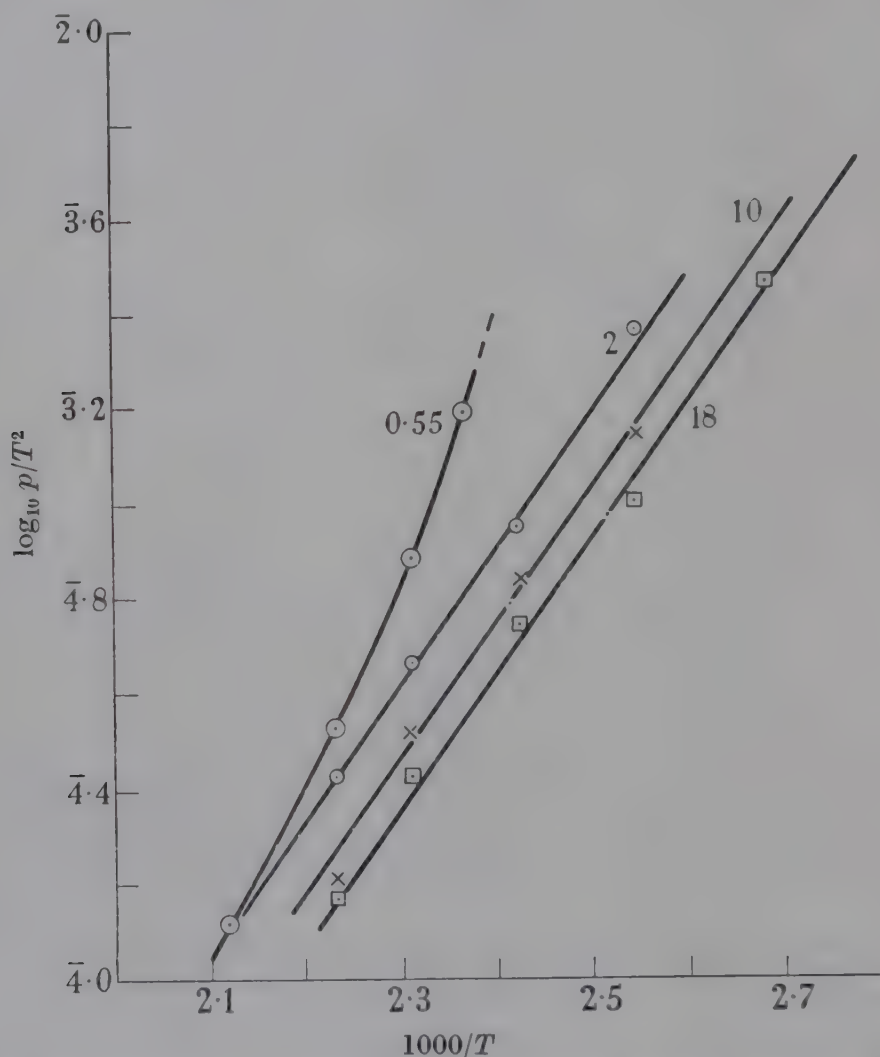


FIGURE 5. The temperature dependence of the minimum ignition pressure ($\text{H}_2 + \text{Cl}_2$) in the presence of 0.55 mm. of chloropicrin in vessel B ($\odot$) and of 2, 10, 18 mm. of chloropicrin in vessel A. Plot of $\log p/T^2$ against $1000/T$.

For the larger concentrations of chloropicrin the lines are reasonably straight and nearly parallel, and suggest that the approximations made in deriving equation (11), though crude, are near the facts. For the low concentration of chloropicrin there is a progressive decrease in the slope as higher temperatures are reached; this is probably because the first term in the denominator of (8) becomes important at lower chloropicrin concentrations.

The slow reaction below the ignition limit at temperatures between 100° and 200° C.

The rate equation adopted for the reactions leading to ignition requires some modification when dealing with the slow reaction below the lower limit; for the time

intervals then concerned allowance must be made for the very powerful chain terminating process k_3 (and a similar process k'_3 involving H atoms) introduced by the nitrosyl chloride formed from the decomposition of the chloropicrin. When these reactions are included, the rate equation can be simplified, by making reasonable assumptions (see appendix A) to

$$\frac{d[\text{HCl}]}{dt} = \frac{2k_1 k_6 k_7 [\text{Cl}_2] [\text{H}_2] [\text{CCl}_3\text{NO}_2]}{k_7 [\text{H}_2] \{k_5 [\text{CCl}_3\text{NO}_2] + k'_3 [\text{NOCl}]\} + k_6 [\text{Cl}_2] \{k_4 [\text{CCl}_3\text{NO}_2] + k_3 [\text{NOCl}]\}} \quad (12)$$

Now at any time t , the concentrations $[\text{CCl}_3\text{NO}_2]_t$ and $[\text{NOCl}]_t$ are given by the equations

$$[\text{CCl}_3\text{NO}_2]_t = [\text{CCl}_3\text{NO}_2]_0 e^{-kt}$$

$$[\text{NOCl}]_t = [\text{CCl}_3\text{NO}_2]_0 (1 - e^{-kt}),$$

where k is the velocity constant for the first order decomposition of chloropicrin. Hence the rate equation becomes

$$\frac{d[\text{HCl}]}{dt} = \frac{2k_1 k_6 k_7 [\text{H}_2] [\text{Cl}_2] e^{-kt}}{k_7 [\text{H}_2] \{k_5 e^{-kt} + k'_3 (1 - e^{-kt})\} + k_6 [\text{Cl}_2] \{k_4 e^{-kt} + k_3 (1 - e^{-kt})\}} \quad (13)$$

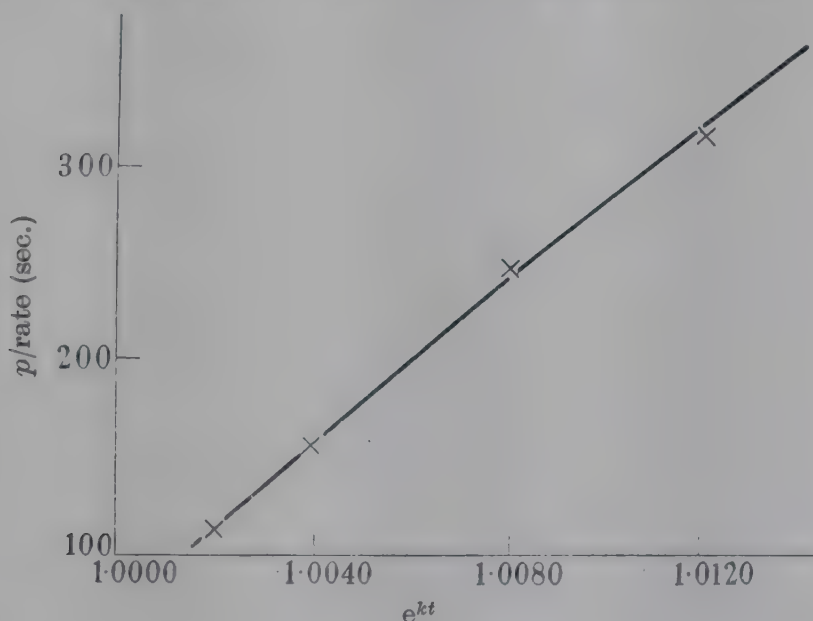


FIGURE 6. Plot of the function p/rate against e^{kt} . p is the pressure (mm.) of chlorine, rate in mm. of chlorine per second, k is the velocity constant for the decomposition of chloropicrin (sec^{-1}), t in sec. 162°C .

The important conclusion may be drawn from this equation that the rate of the slow reaction is independent of the concentration of chloropicrin, especially near the ignition limit where the approximations made in deriving (13) are most valid. This is in agreement with figure 8 of part IX.

When $[\text{H}_2] = [\text{Cl}_2] = p/2$, (13) becomes

$$\frac{d[\text{HCl}]}{dt} = \frac{p e^{-kt}}{F + G e^{-kt}}.$$

Hence

$$\frac{p}{\text{rate}} = G + F e^{kt}. \quad (14)$$

Equation (13) can therefore be tested by plotting p/rate against e^{kt} . At 162°C , the half life of the reaction $\text{CCl}_3\text{NO}_2 \rightarrow \text{COCl}_2 + \text{NOCl}$ is about 1600 sec. (Smith & Steacie) and hence kt is small for $t < 50$ sec. The values of p/rate have been calculated from the smooth curve drawn through the experimental points in figure 8, part IX, and have been plotted against e^{kt} in figure 6. The points are a reasonable fit to a straight line as required by equation (14). This lends support to the expressions derived for the rate equations, but cannot be pressed further in the absence of more experimental data.

CONCLUSIONS

The kinetic study outlined in this paper has shown that the assumption that chlorine atoms are liberated during the decomposition of chloropicrin fits the experimental observations of Smith & Steacie. Estimates of the relative concentration of chlorine atoms at various stages of the decomposition at different temperatures have been made and are in accordance with the results of experiments on the sensitization of ignitions of mixtures of hydrogen and chlorine and of hydrogen and oxygen.

The authors wish to thank the Governing Body of Emmanuel College, Cambridge, for a Research Studentship given in 1938 to one of us (P.G.A.) and for a Research Fellowship (P.G.A.) which enabled this work to be resumed in 1946. They also wish to thank the Department of Scientific and Industrial Research for a Research Maintenance Grant in 1938 (P.G.A.). They also wish to thank Dr F. S. Dainton and Dr W. G. Burns for many helpful discussions.

APPENDIX A

Derivation of the rate equations with straight chains in the hydrogen-chloropicrin reaction.

The kinetic scheme outlined on p. 43 leads to the following stationary state equations involving the chain centres

$$k_1[\text{CCl}_3\text{NO}_2] + k_6[\text{H}][\text{Cl}_2] = A[\text{Cl}] \text{ where } A = \{k_7[\text{H}_2] + k_8 + k_4[\text{CCl}_3\text{NO}_2] + k_3[\text{NOCl}]\}$$

$$\text{and } k_7[\text{Cl}][\text{H}_2] = B[\text{H}] \text{ where } B = \{k_6[\text{Cl}_2] + k_9 + k_5[\text{CCl}_3\text{NO}_2] + k'_3[\text{NOCl}]\}$$

$$\text{therefore } [\text{Cl}] = \frac{k_1[\text{CCl}_3\text{NO}_2] B}{AB - k_6 k_7 [\text{H}_2] [\text{Cl}_2]}$$

$$\text{and } [\text{H}] = \frac{k_1 k_7 [\text{H}_2] [\text{CCl}_3\text{NO}_2]}{AB - k_6 k_7 [\text{H}_2] [\text{Cl}_2]}$$

therefore

$$\frac{d[\text{HCl}]}{dt} = k_7[\text{Cl}][\text{H}_2] + k_6[\text{H}][\text{Cl}_2] = \frac{k_7[\text{H}_2] k_1[\text{CCl}_3\text{NO}_2]}{AB - k_6 k_7 [\text{H}_2] [\text{Cl}_2]} \{B + k_6[\text{Cl}_2]\}.$$

If the chains are long, as they will be near the ignition limit,

$$k_6[\text{Cl}_2] \gg k_9 + k_5[\text{CCl}_3\text{NO}_2] + k'_3[\text{NOCl}]$$

$$k_7[\text{H}_2] \gg k_8 + k_4[\text{CCl}_3\text{NO}_2] + k_3[\text{NOCl}]$$

therefore

$$\frac{d[\text{HCl}]}{dt} = \frac{2k_1[\text{CCl}_3\text{NO}_2] k_6 k_7 [\text{H}_2] [\text{Cl}_2]}{k_6 [\text{Cl}_2] \{k_8 + k_4 [\text{CCl}_3\text{NO}_2] + k_3 [\text{NOCl}]\} + k_7 [\text{H}_2] \{k_9 + k_5 [\text{CCl}_3\text{NO}_2] + k'_3 [\text{NOCl}]\}} \quad (\text{A. 1})$$

It is of importance to estimate when the nitrosyl chloride will begin to have a definite effect upon the rate of chain termination. At 170°C the half life of the main decomposition ($\text{CCl}_3\text{NO}_2 \xrightarrow{k_2} \text{COCl}_2 + \text{NOCl}$) is about 700 sec. (Smith & Steacie 1938). In the short time which elapses before ignition at this temperature (considerably less than 0.5 sec. when the chloropicrin pressure is above 1 mm.) the amount of NOCl formed will be less than 6×10^{-4} mm. per mm. of chloropicrin present. Burns & Dainton (1950) have estimated the activation energy and frequency factor of the reaction $\text{Cl} + \text{NOCl} \rightarrow \text{NO} + \text{Cl}_2$ and at 170°C the velocity constant should be about 3.4×10^{12} ml. mole $^{-1}$ sec. $^{-1}$. The velocity constant of the reaction $\text{Cl} + \text{H}_2 \rightarrow \text{HCl} + \text{H}$ can also be calculated by the collision theory from the known activation energy (Morris & Pease 1935) as about 3×10^{11} ml. mole $^{-1}$ sec. $^{-1}$ at 170°C , so that the rates of these propagation and termination steps, just before ignition in the presence of 1 mm. of chloropicrin and 50 mm., of hydrogen, are in the ratio

$$\frac{3 \times 10^{11} \times 50}{3.4 \times 10^{12} \times 6 \times 10^{-4}} \text{ or } 7 \times 10^3.$$

The chain length of the unsensitized hydrogen-chlorine reaction at these temperatures would probably be greater than 10^4 (Morris & Pease 1939) but might be considerably shortened in the sensitized reaction by the reactions k_4 and k_5 . Evidently the effect of NOCl will not be considerable during the short time before ignition.

Under these circumstances, equation (A. 1) reduces to

$$\frac{d[\text{HCl}]}{dt} = \frac{2k_1 k_6 k_7 [\text{H}_2] [\text{Cl}_2] [\text{CCl}_3\text{NO}_2]}{k_6 [\text{Cl}_2] \{k_8 + k_4 [\text{CCl}_3\text{NO}_2]\} + k_7 [\text{H}_2] \{k_9 + k_5 [\text{CCl}_3\text{NO}_2]\}}$$

and this is the equation (1) which is applied on p. 43 to the sensitized ignitions.

At later times, however, $k_3[\text{NOCl}]$ and $k'_3[\text{NOCl}]$ may become greater than k_8 and k_9 respectively (see p. 43). Equation (A. 1) then reduces to equation (12) p. 47, which is applied to the slow reaction.

APPENDIX B

The isothermal condition for ignition with a branched-chain kinetic scheme.

If the branching is assumed to be due to the reactions



allowance must also be made for the deactivation process



In conjunction with the propagation steps k_6 and k_7 and termination steps k_8 , k_9 , k_4 and k_5 given in the kinetic scheme on p. 43, that is, treating the early stages when $[\text{NOCl}]$ is very small, these equations lead to a net branching factor given by

$$\phi = \frac{k_C k_6 k_7 [\text{H}_2] [\text{Cl}_2] [\text{CCl}_3\text{NO}_2]}{\{k_7 [\text{H}_2] + k_8 + k_4 [\text{CCl}_3\text{NO}_2]\} \{k_D [\text{M}] + k_C [\text{CCl}_3\text{NO}_2]\}} \\ - \frac{k_7 [\text{H}_2] \{k_9 + k_5 [\text{CCl}_3\text{NO}_2]\} + k_6 [\text{Cl}_2] \{k_8 + k_4 [\text{CCl}_3\text{NO}_2]\} + \{k_9 + k_5 [\text{CCl}_3\text{NO}_2]\} \{k_8 + k_4 [\text{CCl}_3\text{NO}_2]\}}{\{k_7 [\text{H}_2] + k_8 + k_4 [\text{CCl}_3\text{NO}_2]\}}.$$

Applying the conditions for ignition $\phi \geq 0$, and making the same assumptions as in appendix A upon the relative values of the rate constants, and also that $k_D \gg k_C$, the condition for ignition is

$$k_C k_6 k_7 [\text{H}_2] [\text{Cl}_2] [\text{CCl}_3\text{NO}_2] \\ = k_D [\text{M}] \{k_7 [\text{H}_2] \{k_9 + k_5 [\text{CCl}_3\text{NO}_2]\} + k_6 [\text{Cl}_2] \{k_8 + k_4 [\text{CCl}_3\text{NO}_2]\}\}.$$

When $[\text{H}_2] = yp$ and $[\text{Cl}_2] = (1-y)p$, this reduces to an expression which is independent of p and suggests that the rate of branching would be independent of pressure and hence that there would be no pressure limit of ignition. This is not observed experimentally, and hence this branching scheme is inadequate.

If the branching is assumed to occur in the three-body reaction



then the isothermal condition for ignition can be shown to be

$$k_C k_7 [\text{H}_2] [\text{Cl}_2] [\text{CCl}_3\text{NO}_2] = k_6 [\text{Cl}_2] \{k_8 + k_4 [\text{CCl}_3\text{NO}_2]\} + k_7 [\text{H}_2] \{k_9 + k_5 [\text{CCl}_3\text{NO}_2]\}.$$

When $[\text{H}_2] = yp$ and $[\text{Cl}_2] = (1-y)p$ this reduces to

$$[\text{CCl}_3\text{NO}_2] \{k_C k_7 y(1-y)p - k_6 k_4(1-y) - k_7 k_5 y\} = \frac{k_7 k_9' y + k_6 k_8'(1-y)}{p}$$

which is of the form $[\text{CCl}_3\text{NO}_2] \{p - B\} = \frac{C}{p}$ when y is constant and is therefore of the same form as equation (5) page 44, and will fit the facts equally well.

REFERENCES

- Ashmore, P. G. & Norrish, R. G. W. 1950a *Proc. Roy. Soc. A*, **203**, 454.
 Ashmore, P. G. & Norrish, R. G. W. 1950b. *Proc. Roy. Soc. A*, **203**, 472.
 Burns, W. G. & Dainton, F. S. 1950 (In the press).
 Dainton, F. S. 1942 *Trans. Faraday Soc.* **38**, 227.
 Dainton, F. S. 1947 *Trans. Faraday Soc.* **43**, 365.
 Dainton, F. S. & Norrish, R. G. W. 1941 *Proc. Roy. Soc. A*, **177**, 411, 421.
 Frank-Kamenetzki, D. A. 1939 *Acta Physicochim. U.S.S.R.* **10**, 365.
 Morris, J. C. & Pease, R. N. 1935 *J. Chem. Phys.* **3**, 796.
 Morris, J. C. & Pease, R. N. 1939 *J. Amer. Chem. Soc.* **61**, 391.
 Semenov, N. 1935 *Chemical kinetics and chain reactions*, p. 41. Oxford University Press.
 Skinner, H. A. 1945 *Trans. Faraday Soc.* **41**, 645.
 Smith, W. McF. & Steacie, E. W. R. 1938a *Canad. J. Res.* **168**, 1, 122.
 Smith, W. McF. & Steacie, E. W. R. 1938b *J. Chem. Phys.* **6**, 145.

Compressibility and absorption frequency of ionic crystals*

BY B. SZIGETI

Department of Theoretical Physics, University of Liverpool

*(Communicated by N. F. Mott, F.R.S.—Received 16 January 1950—
Revised 19 July 1950)*

A simple relation is derived between the compressibility and absorption frequency of ionic crystals. This relation takes into account the change which a displacement induces in the electronic configuration, and is independent of the uncertainty concerning the Lorentz force. Comparison with experimental data shows very good agreement in the case of the alkali halides. The changes in the electronic configuration of each ion are described in terms of a dipole moment and in terms of 'local distortions'. The latter are caused by forces not included in the Lorentz force, and it is assumed that they are localized in small regions in the immediate neighbourhood of the neighbouring ions. The local distortions are responsible for the deviation of the polarization from the value given by the Lorentz force; from the sign of this deviation it is concluded that they are due mainly to the non-electrostatic repulsive forces.

Although this model allows deviations from spherical charge distribution around the ions, it nevertheless permits the splitting of the energy of certain types of displacements into two-body interactions and leads to the Cauchy relations between the elastic constants.

INTRODUCTION

The aim of the present paper is to obtain a relation between the compressibility and absorption frequency of ionic crystals. Expressions for both these quantities have been derived by Born (Born & Göppert-Mayer 1933) in terms of fundamental lattice properties; but in these expressions the deformations of the electron shells have been neglected. It is now known, however, that the electrostatic effect of these deformations is appreciable. This can be calculated in a first approximation, as Born himself pointed out, by developing the charge distribution of each ion into multipoles and maintaining the dipolar term only. This dipolar term gives an important contribution to the dielectric constant, and Lyddane, Sachs & Teller (1941) also showed that it affects appreciably the electrostatic part of the restoring force on which the absorption frequency depends.

But the dipolar approximation would be quite inappropriate for the calculation of the effect of the electronic deformations on the non-electrostatic repulsive forces. These forces are determined by the charge distribution where neighbouring ions overlap: the shift of charge occurring in a displacement need not be uniform over the whole ion, in which case the shift of charge in the overlapping regions and the effect on the repulsive forces bear no relation to the dipole moment associated with the distortions. As the repulsive forces largely determine the compressibility and are also of great importance for the absorption frequency, the principal task will be to find a way to take into account the interaction of the electronic distortions with the repulsive forces. The alkali halides will be primarily in view, as these materials come nearest to extreme ionic behaviour.

* This paper is based on E.R.A. Report L/T 233.

With regard to such distortions, great importance is usually attached to the fact that the Cauchy relation between the elastic constants is fulfilled in the case of the alkali halides. This relation certainly holds if in an elastic deformation the ions interact with central forces, and its validity is therefore usually considered a strong argument in favour of the view that the ions are spherically symmetrical and that in elastic (acoustical) deformations, which include compression, the distortion of the electron shells can be neglected. This is, of course, in agreement with the assumption that all electronic distortions are adequately expressed by a dipolar term, because in elastic deformations a dipole moment cannot arise. It will be shown, however, that electronic distortions in an elastic deformation are not incompatible with central forces.

It will be shown in the present paper that if we use a dipolar term only and assume that this is determined by the Lorentz force, then a simple relation can be derived between the compressibility and absorption frequency which is fulfilled for alkali halides. But the dipolar approximation leads to a value of the dielectric constant which does not agree well with experiment. Several authors (Mott & Gurney 1948) suggested that this discrepancy is due to the spatial extension of the dipoles and, while maintaining the dipolar approximation, introduced a correction to the Lorentz force so as to obtain the right dielectric constant; but if this is done, then the relation between the compressibility and absorption frequency has to be modified and no longer agrees with experiment. It thus seems that the dipolar approximation is not adequate if we want both the relation between the absorption frequency and compressibility and also the expression for the dielectric constant to agree with experiment.

In any case, the failure of the Lorentz force is an indication that complicated distortions must occur. This may be seen from Mott & Gurney's discussion (1948) of these problems, who concluded that in order to obtain a correction to the Lorentz force of the required magnitude, it is necessary to assume that neighbouring ions penetrate considerably into each other and that the electrostatic field exerted in the overlapping regions is very different from its value at the ionic centres. The correction term to the Lorentz force then represents a force of very short range, acting mainly in the overlapping regions, and the shift of charge it produces is likely to be mainly concentrated in the same regions and therefore to have a strong effect on the repulsive interaction. Also, if the deviation from the Lorentz approximation is due to the forces exerted in the overlapping regions, then it may be that most, or at least part, of this deviation is caused by forces which are not electrostatic in nature.

With the help of a few simple assumptions, it is proposed to construct a model which takes account in a rough way of the electronic distortions which are not represented adequately by a dipolar term. Although this model allows deviations from spherical charge distribution, it nevertheless leads to the Cauchy relation and also to the relation between the compressibility and absorption frequency which we have mentioned, while at the same time it explains the observed deviations from the Lorentz force.

LONG-RANGE AND SHORT-RANGE FORCES

If δW is the change of the internal energy of the crystal in an arbitrary displacement, it will be convenient to divide this into two parts

$$\delta W = \delta V + \delta \Phi, \quad (1)$$

where V is that part of the electrostatic interaction energy which is obtained if the charge and polarization of each ion is replaced by a point charge and a point dipole placed in its centre, i.e. if poles of higher order and the spatial extension of the ions are neglected. $\delta \Phi$ contains all the rest of the energy change: that part of the electrostatic interaction which is not included in δV , the van der Waals and the repulsive energies, and also the homopolar interaction energy, if any. In addition, not only the interaction between ions but the self-energy of each ion has also been altered due to the distortion of its electron cloud, and this effect is also contained in $\delta \Phi$.

It will be assumed that if the bonds have any homopolar character, this is not strong enough to lead to appreciable directional forces, though it may cause some deviations from spherical charge distribution. When dividing the crystal into individual ions, the valency electrons will be considered part of the negative ions.

V will be called the long-range and ϕ the short-range interaction, as the contribution of the interaction of any two ions to ϕ falls off much more rapidly with their distance than their contribution to V . ϕ may contain contributions due to forces of intermediate range, arising from that part of the electrostatic interaction which is not included in V ; but these contributions are presumably small compared with those due to forces of very short range. In any case, we are interested in the energy change which occurs in a displacement, and this change depends much more strongly on the rapidly varying forces than does the energy itself. It will be therefore sufficient to include in $\delta \phi$ the change of the self-energy of the ions, and the interaction between nearest and, to a much smaller extent, between second nearest neighbours.

In contrast to the interaction of intermediate range, δV may give an important contribution to the energy change, because this interaction falls off so slowly with distance that in the contribution of any particular ion its interaction with very many others has to be considered. δV is particularly large in the case of polarization and is responsible for the considerable difference between the frequencies of longitudinal and transverse polarization waves.

In addition to the internal forces arising out of δV and $\delta \phi$, there may be also an external electrical field acting. In the Lorentz approximation the effective field F acting on an ion is calculated by replacing the charge and polarization of each ion by a point charge and point dipole placed at its centre. The Lorentz effective field thus represents exactly the forces due to δV and to any electrical field which may be exerted by external sources. For cubic crystals it is given by the well-known formula

$$F = \frac{4\pi}{3} P + E, \quad (2)$$

where P is the polarization and E the macroscopic electric field. Any additional forces are due to $\delta \phi$ and are thus mainly forces of very short range, acting where

nearest and, to a much smaller extent, where second nearest neighbours overlap. An entirely different argument in the introduction also led to this conclusion. It will be therefore assumed that, in a first approximation, the electronic deformations are due to two causes:

(a) an electrostatic field, expressed by the Lorentz field F , resulting in a shift of charge over the whole ion;

(b) forces which act only in the regions where neighbouring ions touch or overlap and cause local distortions in these regions.

THE LOCAL DISTORTIONS

We shall start by discussing those displacements where $F = 0$, so that only local distortions due to short-range forces occur. This includes the following three cases:

(1) Isotropic compression or expansion.

(2) Long elastic (acoustical) waves, in those crystal structures where each ion is a centre of symmetry. This includes the NaCl and CsCl, but not the ZnS structures.

(3) Uniform polarization of a cubic crystal of spherical shape, in absence of an external electric field.

In the first two cases both E and P are zero, while in the polarized sphere the only electric field is due to the surface charges and is given by $E = -4\pi P/3$. Equation (2) thus shows that F vanishes in all three cases.

The short-range forces act mainly in the regions where neighbouring ions touch or overlap. Although the distortions they produce presumably extend farther than the range of these forces, it will nevertheless be assumed that the distortions induced on an ion by its various neighbours do not extend very far and are independent of each other. Let $\delta\phi_{ij}$ represent the change of the short-range interaction between neighbouring ions i and j and also the energy of the distortions which these two ions induce on each other, and δr_{ij} the change of distance between them. Then $\delta\phi_{ij}$ is a function of δr_{ij} and of the distortions of the two ions in the neighbourhood of each other. Since these distortions, in turn, only depend on the interaction of the two ions and are independent of forces exerted by other ions, $\delta\phi_{ij}$ is a unique function of δr_{ij} . $\delta\phi$ can thus be split into two-body interactions

$$\delta\phi = \sum_{i,j} \delta\phi_{ij}(\delta r_{ij}) \quad (3)$$

if $F = 0$. Only such ion pairs are included in the summation which are first or second neighbours, as only these contribute to $\delta\phi$.

Figures 1 and 2 show the crystal in equilibrium and in compression, respectively. In order to simplify matters, only the distortions of the negative ion by nearest neighbours are shown, as these distortions are presumably the most important. The bulges facing nearest neighbours are intended to convey that the charge density in those regions is higher than would correspond to spherical symmetry. The figures are not intended to indicate that the ions do not overlap; in fact, in view of what has been said, most of the distortion takes place where the overlap is appreciable.

Most probably there is some deviation from spherical symmetry even in equilibrium, due to the short-range interaction with neighbours. The figures illustrate

the case when in equilibrium there is a bulge which is, however, rapidly diminished as the neighbours approach closer. As a small homopolar tendency primarily results in the accumulation of charge in the regions where the two ions overlap, the figures

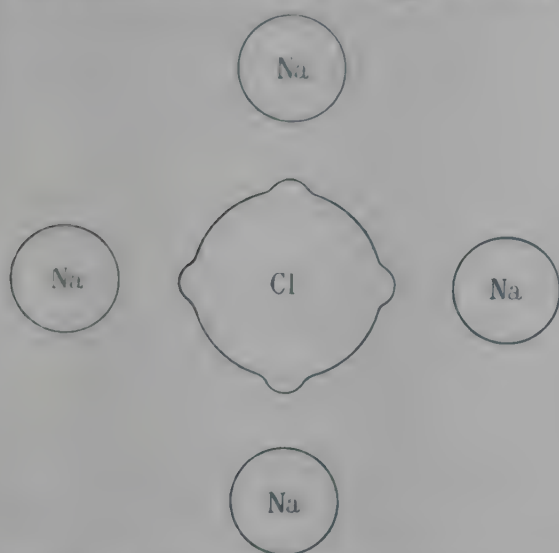


FIGURE 1. Ions in their equilibrium positions. Only the distortion of the negative ion due to nearest neighbours is shown.

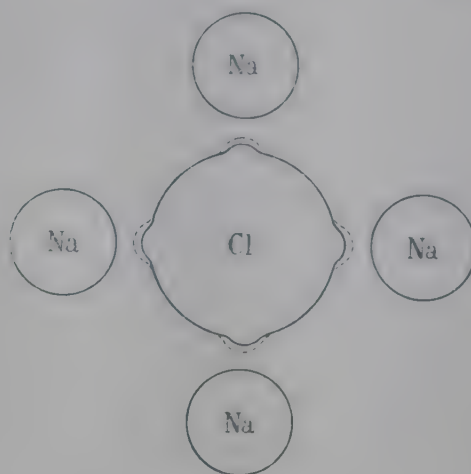


FIGURE 2. Compression. The dashed line shows the shape of the electron cloud in figure 1, i.e. before the crystal was compressed.

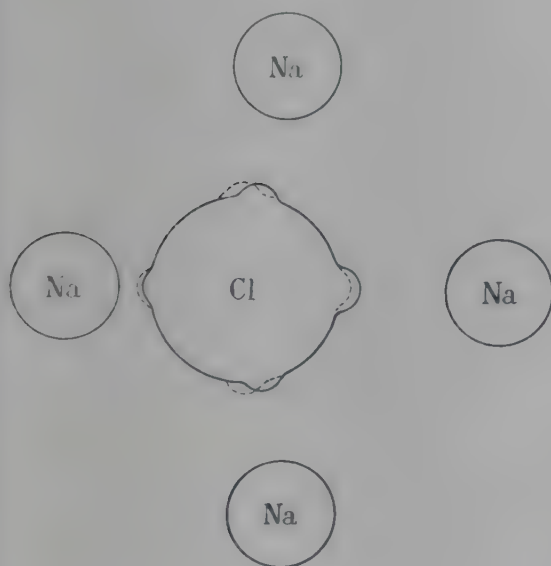


FIGURE 3. Spontaneous polarization of the lattice if $F = 0$. The electronic contribution to this particular type of polarization consists only in the change of the local distortions. The dashed line again shows the shape of the electron cloud in figure 1, i.e. before the ions were displaced from their equilibrium positions.

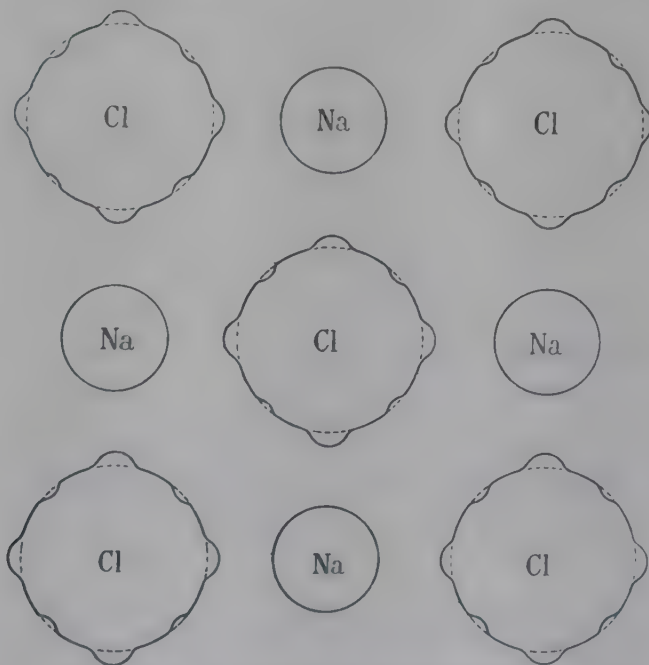


FIGURE 4. Ions in their equilibrium positions. In this figure the distortions of the negative ions due to second neighbours are also indicated. The dotted line corresponds to spherical shape, for the sake of comparison.

may be taken to illustrate the case when there is a small homopolar tendency in equilibrium which diminishes with decreasing distance between neighbouring ions. Possible distortions of the negative ions due to second neighbours, neglected in figures 1 to 3, are shown in figure 4.

Equation (3) is formally the same as would be obtained for spherical, non-deformable ions, even though $\delta\phi_{ij}$ contains the effect of the electronic deformations. The assumption which led to this result was that the short-range forces from the various neighbours cause localized distortions which are independent of each other. This may be only a rough approximation, but should nevertheless be nearer to the true state of affairs than either neglecting these distortions altogether, or expressing them by a dipolar term representing a uniform shift of charge over the whole ion.

The situation is more complicated in displacements where F does not vanish, because the electron shift then extends over the whole ion and the charge distribution in the region where two ions touch is affected by forces exerted by other ions. In that case, it is not possible to break up $\delta\phi$ into two-body interactions and equation (3) does not hold. There is, nevertheless, no need to construct a model for this more general case, as there are rigorous macroscopic relations (Lyddane *et al.* 1941; Szigeti 1949) connecting the frequency, polarizability and effective charge for the uniform polarization of a sphere, where $F = 0$, with the same quantities for other cases, where $F \neq 0$, such as longitudinal or transverse optical waves.

POLARIZATION

The ions in figure 3 represent that case of polarization where F vanishes and only local distortions occur, viz. the uniformly polarized sphere in absence of an external electric field. Again, only the distortions of the negative ion are shown in the figure; these depend on nearest neighbour distance in the same way as in compression. One bulge increases and the other diminishes by the same amount, because one neighbour comes closer and the other moves farther away. The other two bulges, as well as those two which are not in the plane of the figure, do not change their magnitude, since the distance of these neighbours is not changed in the first order; these four bulges, however, change their directions slightly since each bulge always faces the ion which is adjacent to it.

The magnitude of the electronic polarization illustrated in figure 3 may be obtained from a recently derived relation (Szigeti 1949):

$$\epsilon - n^2 = \left(\frac{n^2 + 2}{3}\right)^2 4\pi N \frac{(sze)^2}{m\omega_i^2}, \quad (4)$$

where ϵ is the static dielectric constant, n the optical refractive index, ω_i the infra-red absorption frequency and m , the reduced mass of an ion pair, is defined by

$$\frac{1}{m} = \frac{1}{m_1} + \frac{1}{m_2}, \quad (5)$$

m_1 and m_2 being the masses of the two atoms. z is the valency of the ions and e the electronic charge. s is defined by $\mu_s(x) = szez$,

where μ_s is the dipole moment per ion pair in a spontaneously polarized sphere if x is the displacement of the two kinds of ions relative to each other. As F vanishes in the Lorentz approximation, there should be no electronic polarization in this case, and μ_s ought to be equal to zez . In other words, in the Lorentz approximation s would be unity.

On the other hand, s may be calculated from equation (4), as all other quantities in that equation can be determined by experiment. Values of s so obtained are reproduced from the previous paper (Szigeti 1949) in the first column of table 1. It was explained in the previous paper that the figures for the alkali halides are believed to be fairly accurate, but those for some of the other materials may be in error due to the uncertainty of the experimental technique used.

TABLE 1

Substance	lattice type	s	R (10^{-8} cm.)	$R^2/3u$ (10^6 cm.)	ϵ	n^2	$m\omega_i^2$ (10^4 dynes/cm.)	$\kappa_{\text{calc.}}$ (10^{-6} cm. ² kg.)	$\kappa_{\text{obs.}}$ (10^{-6} cm. ² kg.)	$\kappa_{\text{calc.}}, \kappa_{\text{obs.}}$
LiF	NaCl	0.87	2.07	8.05	9.27	1.92	2.8	1.5	1.5	1.0
NaF	NaCl	0.93	2.31	7.22	6.0	1.74	3.7	1.72	2.07	0.83
NaCl	NaCl	0.74	2.81	5.94	5.62	2.25	2.2	4.13	4.18	0.99
NaBr	NaCl	0.69	2.97	5.61	5.99	2.62	1.8	5.62	4.98	1.13
NaI	NaCl	0.71	3.23	5.16	6.60	2.91	1.5	7.3	6.94	1.05
KCl	NaCl	0.80	3.14	5.31	4.68	2.13	2.2	5.28	5.53	0.96
KBr	NaCl	0.76	3.29	5.07	4.78	2.33	2.0	6.23	6.56	0.95
KI	NaCl	0.69	3.53	4.72	4.94	2.69	1.7	8.3	8.37	0.99
RbCl	NaCl	0.84	3.27	5.1	5	2.19	2.0	5.82	6.52	0.89
RbBr	NaCl	0.82	3.42	4.88	5	2.33	1.9	6.48	7.78	0.83
RbI	NaCl	0.89	3.66	4.56	5	2.63	2.3	6.17	9.39	0.66
CsCl	CsCl	0.84	3.56	6.08	7.20	2.60	1.6	5.05	5.83	0.87
CsBr	CsCl	0.79	3.71	5.84	6.51	2.78	1.6	6.0	6.92	0.87
TlCl	CsCl	1.08	3.33	6.52	31.9	5.10	1.3	2.44	4.8	0.51
CuCl	ZnS	1.10	2.34	4.63	10	3.57	4.7	2.1	2.46	0.85
CuBr	ZnS	0.995	2.46	4.40	8	4.08	6.5	2.08	2.87	0.72
MgO	NaCl	0.88	2.10	7.94	9.8	2.95	18.8	0.28	0.59	0.47
CaO	NaCl	0.76	2.40	6.94	11.8	3.28	8.9	0.60	(4.57)	(0.13)
SrO	NaCl	0.58	2.57	6.48	13.3	3.31	3.6	1.46	—	—
ZnS	ZnS	0.48	2.33	4.64	8.3	5.07	11.6	1.25	1.28	0.98

It is seen that for most materials, including the alkali halides, s differs from unity considerably. It is perhaps worth stressing that the reason for this deviation cannot lie in the fact that equation (4) is based on a first-order approximation in the displacement, because if this were the case then the discrepancy should disappear at low temperatures. But not only is the observed temperature variation of the dielectric constant much smaller than one would expect in this case, but in many cases, including all those alkali halides for which it has been measured, it is even of the wrong sign.

On the basis of our model, the deviation of s from unity is due to the electronic dipole moment created by the local distortions. Equation (6) shows that this dipole moment μ' is given by

$$\mu' = (s - 1) z e x. \quad (7)$$

The distortions of the positive ions will be neglected in the following qualitative considerations, as in figure 3. In the case of the alkali halides $s \approx 0.8$, and a simple calculation shows that those four bulges which only change direction but not magnitude can hardly account for the dipole moment required in equation (7), as this would imply a much larger deviation from spherical symmetry in the equilibrium position than appears likely. Those two bulges, however, which change their magnitude may easily provide the required shift of charge, since their contribution does not depend on the original magnitude of the bulge, but on its *derivative* with respect to nearest neighbour distance. This derivative is likely to be much larger than the original

bulge, as the derivative of the short-range forces, which cause the bulge, is much larger than the short-range forces themselves.

In view of this, equation (7) gives no indication and it is immaterial for the present argument, whether the distortion in equilibrium is a bulge, as in figure 1, or a dent. But as the change of the bulge is mainly responsible for the dipole moment (7), and as $(s-1)$ is negative, the sign of this *change* must be as indicated in the figures, i.e. as the neighbouring ion approaches the electron cloud is forced back. This can certainly not be caused by the electrostatic part of $\delta\phi$ because, as neighbouring ions approach one another and penetration increases, the electrostatic forces tend to increase the bulge; only the non-electrostatic repulsive forces can cause the electron cloud to move back. We thus conclude that in alkali halides the local distortions are governed mainly by the quantum-mechanical repulsive forces.

THE ELASTIC CONSTANTS

In the NaCl and CsCl structures, which include all alkali halides, in long elastic waves $F = 0$ and equation (3) holds. Further, since each ion remains symmetry centre, an electronic dipole moment cannot arise and therefore

$$V_{ij} = \frac{z_i z_j e^2}{z_{ij}}, \quad (8)$$

where z_i and z_j are the valencies. Hence, in view of equation (1), δW_{ij} is a unique function of δr_{ij} . The expression for the energy change in elastic waves is thus formally the same as if the ions interacted with central forces. It is well known that if this is so then the Cauchy relation between the elastic constants holds. Using Voigt's notation, for cubic crystals this relation is

$$c_{12} = c_{44}. \quad (9)$$

Our model thus leads to the Cauchy relation, notwithstanding that it allows deviations from spherical symmetry. It is known from experiment (Durand 1936; Seitz 1940) that equation (9) holds for alkali halides but not for MgO. From this we conclude that our model represents a good approximation for the alkali halides, but not for MgO.

COMPRESSIBILITY AND ABSORPTION FREQUENCY

While second neighbour interaction is important in connexion with the elastic constants c_{12} and c_{44} , in connexion with the compressibility and absorption frequency it is not. In order to be able to carry out further calculations, in this section we shall neglect the short-range interaction between second neighbours. Denoting the total energy of an ion pair and its two parts due to the long-range and short-range interaction by w , v and ϕ , respectively, and the nearest neighbour distance by R , we have

$$\left(\frac{dw}{dR}\right)_c = \left(\frac{dv}{dR}\right)_c + \left(\frac{d\phi}{dR}\right)_c, \quad (10)$$

where the subscript c indicates that R has to be changed without change in crystal structure, i.e. the derivative has to be taken with respect to uniform compression or expansion. The expression is zero, because the crystal is in equilibrium. Equation (10) as well as most of the calculations in this section follow well-established lines,

except that instead of separating the energy into electrostatic and repulsive parts in the usual manner, our v and ϕ are defined in a somewhat different way.

In uniform compression or expansion no electronic dipole moment is induced and v is the same as it would be for unpolarizable ions. Therefore $(dv/dR)_c = M(ze)^2/R^2$, where M , the Madelung constant, depends on the crystal structure only. Let ϕ' and ϕ'' denote the first and second derivatives of the short-range interaction between two neighbouring ions and ζ the number of nearest neighbours. Equation (10) then becomes

$$\left(\frac{dw}{dR}\right)_c = M \frac{(ze)^2}{R^2} + \zeta\phi' = 0. \quad (11)$$

And for the second derivative we have

$$\left(\frac{d^2w}{dR^2}\right)_c = -2M \frac{(ze)^2}{R^3} + \zeta\phi''. \quad (12)$$

In a crystal the following equation holds for the compressibility κ :

$$\frac{1}{\kappa} = u \left(\frac{d^2w}{du^2} \right)_c,$$

where u is the volume occupied by an ion pair. This equation can be written

$$\frac{1}{\kappa} = u \left(\frac{d^2w}{dR^2} \right)_c \left(\frac{dR}{du} \right)_c^2,$$

because $(dw/dR)_c = 0$. In compression u varies as R^3 , so that $(du/dR)_c = 3u/R$ and, using equation (12),

$$\frac{1}{\kappa} = \frac{R^2}{9u} \left(\zeta\phi'' - 2M \frac{(ze)^2}{R^3} \right). \quad (13)$$

Next we consider a sphere which is uniformly polarized in the absence of an external electrical field. The energy change is

$$\delta w = \frac{1}{2} m \omega_s^2 x^2, \quad (14)$$

if ω_s is the frequency for this particular type of displacement, m the reduced mass given by equation (5) and x the displacement of the two kinds of ions relative to each other. Equation (8) is now not valid because the displacement is asymmetric and, as figure 3 shows, an electronic dipole moment is induced which contributes to the V_{ij} 's; on the other hand, the sum of the long-range interactions of any particular ion with all others vanishes in this case:

$$\sum_j \delta V_{ij} = 0. \quad (15)$$

This follows from the same argument as the conclusion that in a uniformly polarized sphere F is zero (Fowler 1936; Lyddane & Herzfeld 1938). Therefore $\delta v = 0$ and

$$\delta w = \delta \phi. \quad (16)$$

To obtain $\delta \phi$ we have to add up the interaction of one of the ions with its nearest neighbours. Choosing the i th ion, expanding and neglecting terms of higher order than the second, we have

$$\delta \phi = \sum_j \delta \phi_{ij} = \sum_j (\phi' \delta r_{ij} + \frac{1}{2} \phi'' \delta r_{ij}^2), \quad (17)$$

where the summation extends over the nearest neighbours of the i th ion. If in figure 3 the central ion is the i th, then it is seen that δr for one of the neighbours is

+ x and for another $-x$. The combined contribution of these two neighbours to (17) is thus $\frac{1}{2}\phi''x^2$. For each of the other four neighbours a power series development gives $\delta r = x^2/2R$; neglecting terms of higher order than x^2 , the contribution of these four ions to (17) is $\phi'x^2/R$. The total change of energy is therefore

$$\delta\phi = \left(\phi'' + \frac{2\phi'}{R}\right)x^2.$$

Since figure 3 represents the NaCl structure, where $\zeta = 6$, this can also be written

$$\delta\phi = \frac{\zeta}{6} \left(\phi'' + \frac{2\phi'}{R}\right)x^2. \quad (18)$$

It is readily verified that this last equation is also valid for the CsCl and ZnS structures, with the appropriate values of ζ . Combining (18) with (14) and (16) gives

$$m\omega_s^2 = \frac{\zeta}{3} \left(\phi'' + \frac{2\phi'}{R}\right). \quad (19)$$

From (11) we have

$$\phi' = -M \frac{(ze)^2}{\zeta R^2}.$$

Inserting this in equation (19) and comparing with (13) gives

$$\frac{1}{\kappa} = \frac{R^2}{3u} m\omega_s^2. \quad (20)$$

The absorption frequency ω_t , which is the frequency of transverse polarization waves, where $F \neq 0$, is connected to ω_s by a macroscopic relation (Lyddane *et al.* 1941), viz.

$$\omega_t^2 = \frac{n^2 + 2}{\epsilon + 2} \omega_s^2. \quad (21)$$

With (20) this gives the required relation between the compressibility and absorption frequency:

$$\frac{1}{\kappa} = \frac{R^2}{3u} \frac{\epsilon + 2}{n^2 + 2} m\omega_t^2. \quad (22)$$

All the quantities in this equation can be determined by experiment.

COMPARISON WITH EXPERIMENT AND DISCUSSION

Measured values of the quantities in equation (22) have been collected in table 1 for those diatomic crystals for which all the data are available. ω_t has been taken from absorption measurements; the sources for ω_t , ϵ and n have been given in the previous paper (1949). The values for R , u , ϵ , n and $\kappa_{\text{obs.}}$ are the same as used by Höjendahl (1938). $\kappa_{\text{calc.}}$ is the value calculated from equation (22). In the case of CaO, the bracketed figures are very doubtful, as is seen from the original paper by Bridgman (1932) which describes the measurement of the compressibility.

In view of the approximations made and of the margin of experimental error, the agreement for the alkali halides seems very satisfactory. For some of the other materials the agreement is much worse; this indicates that our model is not

applicable to these materials. Presumably not much importance is to be attached to the remarkable agreement in the case of ZnS, which substance is certainly less ionic than the alkali halides. It has been mentioned that the experimental data for some of the crystals which are not alkali halides may contain fairly large errors, and this may well be the case for ZnS.

Our expressions for the energy of displacement when $F = 0$, namely, equation (3) for $\delta\phi$ and equations (8) and (15), respectively, for δV , are formally the same as would be the case if the ions were spherical. Since in the Lorentz approximation the ions are spherical whenever $F = 0$, that approximation also leads to these three equations; while in a transverse wave, where $F \neq 0$, the effect of F is given by the macroscopic equation (21) independently of the nature of the electronic distortions. It is true that the Lorentz model also leads to equation (8) in the case of the uniformly polarized sphere while ours does not, but this makes no difference in the relation between ω_l and κ because of equation (15). Equation (22) connecting ω_l and κ can therefore also be obtained on the basis of the Lorentz approximation, although it has not been derived before.

However, the failure of the Lorentz approximation to give the dielectric constant correctly points to the importance of the non-dipolar distortions. Our model, by taking these into account, reconciles the validity of the Cauchy relation and of equation (22) with the deviation of the dielectric constant from its value given by the Lorentz force.

Nevertheless, all the difficulties cannot of course be solved by such a simple model. While s , representing the electronic polarization associated with the local distortions, has been eliminated from the relation between the compressibility and the absorption frequency, this factor still remains in equation (4) and our model does not make its calculation possible from other fundamental lattice properties. This could presumably be only achieved if the electronic wave functions were calculated both in equilibrium and after displacement; the accuracy required from such calculations would be very exacting.

Returning to those ionic crystals which are not alkali halides, these are believed to have a somewhat stronger tendency than alkali halides to share the valency electrons between neighbouring ions. The charge which each ion carries is in this case somewhat less than ze , and our expression for δV is therefore not accurate. This, however, should not have much effect on the results, since the elastic constants, the compressibility and ω_s depend principally on the short-range forces, while in the case of ω_l , where the contributions of the short-range and the long-range interactions are of about equal importance, the effect of δV has been expressed in terms of ϵ and n^2 in a way which is independent of the charge carried by the ions.

Much more important is the circumstance that as a result of increasing electron sharing the distortions caused by the neighbouring ions are no longer sufficiently localized to be independent of each other. $\delta\phi$ cannot then be separated into two-body interactions and equation (3) is not valid. This must be the reason for the discrepancies between $\kappa_{\text{obs.}}$ and $\kappa_{\text{calc.}}$ in the case of materials which are not alkali halides.

One may expect that one of the consequences of increasing sharing of electrons should be the appearance of directed forces. As the valency angle is bent in polariza-

tion but not in compression, directed forces would cause $\kappa_{\text{calc.}}$ to be smaller than $\kappa_{\text{obs.}}$. The table shows that this is, in fact, the case for all those materials for which the discrepancy between calculated and observed values is large.

The author is indebted for support from the British Electrical and Allied Industries Research Association, and for their permission to publish this paper.

REFERENCES

- Born, M. & Göppert-Mayer, M. 1933 *Handb. Phys.* **24** (2), 623.
 Bridgman, P. W. 1932 *Proc. Amer. Acad. Arts Sci.* **67**, 345.
 Durand, M. A. 1936 *Phys. Rev.* **50**, 449.
 Fowler, R. H. 1936 *Statistical mechanics*, 2nd ed. p. 441. Cambridge University Press.
 Höjendahl, K. 1938 *K. danske vidensk. Selskab.* **16**, no. 2.
 Lyddane, R. H. & Herzfeld, K. F. 1938 *Phys. Rev.* **54**, 846.
 Lyddane, R. H., Sachs, R. G. & Teller, E. 1941 *Phys. Rev.* **59**, 673.
 Mott, N. F. & Gurney, R. W. 1948 *Electronic processes in ionic crystals*, 2nd ed. p. 17. Oxford: Clarendon Press.
 Seitz, F. 1940 *Modern theory of solids*, pp. 94–96. New York: McGraw-Hill Book Co.
 Szigeti, B. 1949 *Trans. Faraday Soc.* **45**, 155.

Kinetic studies on carbohydrates in alkaline conditions.

I. The kinetics of the autoxidation of glucose

BY C. H. BAMFORD AND J. R. COLLINS

Courtaulds Ltd., Maidenhead, Berks

(Communicated by A. H. Wilson, F.R.S.—Received 3 May 1950—
 Revised 15 July 1950)

The autoxidation of glucose and fructose in alkaline solution has been shown to result in the almost quantitative formation of formic and arabonic acids at high oxygen pressures. Under these conditions, the degradation and interconversion reactions are suppressed and the autoxidation reaction reaches a limiting rate.

From a general kinetic examination of the autoxidation reaction, evidence is obtained in support of the idea that, at high oxygen pressures, the rate-determining step is that of the parallel formation of two ionic intermediates. The kinetics of the autoxidation can be completely explained in terms of simple reactions of these intermediates. The absolute velocity constants for the formation of the intermediates have been determined from steady-state experiments. Experiments in the non-steady state have enabled the calculation of the concentrations of the intermediates to be made, and thus the velocity constants for their reaction with oxygen can be estimated.

GENERAL INTRODUCTION

The reactions of simple sugars in alkaline solution are of considerable chemical and biological interest and have been studied for over half a century. Apart from extensive kinetic studies on mutarotation, most of the work has been done on the alkaline degradation and autoxidation of glucose and is of a semiquantitative or

preparative nature. So far as we are aware there have been no previous kinetic studies on the autoxidation or alkaline rearrangements of any hexose from which it would be possible to formulate a reaction scheme. A knowledge of the mechanism of the autoxidation of glucose and its derivatives, e.g. methyl glucoside, in alkaline solution was desirable for at least two reasons. First, it seemed likely that it would be of help in understanding the technically important processes occurring when alkali cellulose is exposed to oxygen. Secondly, it would contribute to our understanding of oxidation processes in general, many of which have been investigated recently and have been shown to proceed through free radical chain mechanisms (see e.g. Bolland & Gee 1946).

Since sugars undergo many changes in alkaline solution at a rate comparable to that of the oxidation, the latter cannot be studied independently. It is therefore desirable to summarize here the relevant information available in the literature about these reactions.

The mutarotation of sugars is believed to occur by an acid-base catalyzed proton interchange, with a rate-determining step resulting in the formation of the open aldehyde form of the sugar, which rapidly reverts to the α - or β -ring form (see Hammett 1940). The important point for present purposes is that the rate of mutarotation of glucose and fructose is very much greater than that of any of the reactions considered in these papers under comparable conditions. It is clear therefore that these reactions contain other rate-determining steps.

The autoxidation of glucose has been shown by Spengler & Pfannenstiel (1935) and Isbell (1942) to give arabonic acid in yields up to 75 % of theory. The other main product was stated to be formic acid, and it is implied in Spengler & Pfannenstiel's paper that fructose gives the same oxidation products as glucose. Earlier workers (Nef 1910, 1914; Glattfeld 1913; Power & Upson 1926) obtained in addition numerous other products. It seems clear that these were not primary products of the oxidation of glucose. These investigators generally used low concentrations of alkali and oxygen, under which conditions degradation and resin formation, together with oxidation of the resulting products, occurred simultaneously.

In the absence of oxygen, interconversion and degradation of the hexoses occur. Glucose, fructose and mannose are mutually interconvertible in alkaline solution, and in very dilute alkali (where the rate of degradation is negligible) an equilibrium tends to be established. The amount of mannose present is small (see, for example, Wolfrom & Lewis 1928); in the present series of papers the formation of this sugar will not be taken into account. It is unlikely that this will lead to appreciable errors, since the rates of oxidation and degradation of mannose would be expected to be very similar to those of glucose.

The course of the degradation reaction is dependent on the concentration of the alkali. If the normality of the alkali is greater than 0.5, both glucose and fructose readily react to produce lactic and saccharinic acids as the main products, together with small amounts of dihydroxybutyric acid and a resin which causes a darkening of the solution. As the normality of alkali is decreased the quantity of resin is increased, and volatile acids, e.g. formic and acetic, are formed. The results of Shaffer & Friedemann (1930) are consistent with an almost quantitative conversion

of the sugar into lactic and saccharinic acids, for hydroxyl concentrations in the range we have studied (0.4 to 4.5 N). Below 0.4 N the amounts of volatile acids and resin become a significant fraction of the total yield.

We have already indicated that there do not appear to be any kinetic measurements on which a mechanism for these reactions can be based. The formation of the products has been discussed mainly in terms of the enediol theory originally formulated by Nef, and developed by Evans and his co-workers (see review by Evans 1929). The original version of the theory assumed that 1,2-, 2,3- and 3,4-enediols were present in alkaline solutions of hexoses and that these split at the double bonds, the resulting fragments either rearranging to give the degradation products, or reacting with oxidizing reagents in the solution. Such a mechanism is clearly incompatible with modern ideas, and Evans (1942) proposed a modification of it in which the splitting of the enediols was supposed to occur at the α - β link. While this is more acceptable, Evans's suggestion leaves the details of the mechanism quite obscure. A mechanism for the formation of saccharinic acids from glucose in modern terms is given by Isbell (1944), who does not, however, consider lactic acid formation.

The present work is a study of the kinetics of the autoxidation, interconversion and degradation reactions of glucose, and has as its object the development of a single scheme incorporating all these processes in a quantitative manner. Investigations of the autoxidation were carried out in steady and non-steady state experiments, described in part I. A combination of the results of these experiments with those on the interconversion and degradation reactions described in part II enables a complete reaction scheme to be drawn up in which the velocity constants of most of the reactions can be calculated absolutely.

A. STEADY-STATE EXPERIMENTS

Experimental

The course of the oxidation was followed by measuring the volume of oxygen absorbed at constant pressure by glucose in sodium hydroxide solution. The concentrations of sodium hydroxide used lay in the range 0.05 to 6.22 N. In order to minimize the effect of the changing medium on the reaction velocity, the concentration of electrolyte was made up to 6.22 N in every case by addition of sodium bromide. This salt was chosen because its activity coefficients are very close to those of sodium hydroxide at similar concentrations. At the lowest alkali concentration addition of the sodium bromide increased the rate by about 30 %.

(i) *Apparatus*

The absorption apparatus was in principle a modified Warburg respirometer (see Dixon 1943), though it differed greatly in construction. An extended diagram of it is given in figure 1.

A dibutyl phthalate manometer *A* was connected between the reaction bulb *B* and the compensation bulb *C*, each of about 60 ml. capacity. The reaction bulb was attached by means of a glass spiral and could be shaken mechanically with frequencies up to 450 per min. with an amplitude of 1.5 to 2 cm. The by-pass tap

D and outlet tap *E* were used for filling the apparatus with oxygen. Oxygen uptake was measured by one of the gas burettes *F* connected to a mercury reservoir *G*. The burettes were of 5, 10 and 25 ml. capacity, so enabling a suitable range to be selected for any experiment. The mercury level was adjusted by altering the air pressure over *G* which was connected to a pressure-vacuum system capable of fine adjustment. In this way the pressures in *B* and *C* were maintained constant, using the manometer *A* as a null instrument. The volume of gas absorbed was thus measured directly by the gas burette in use.

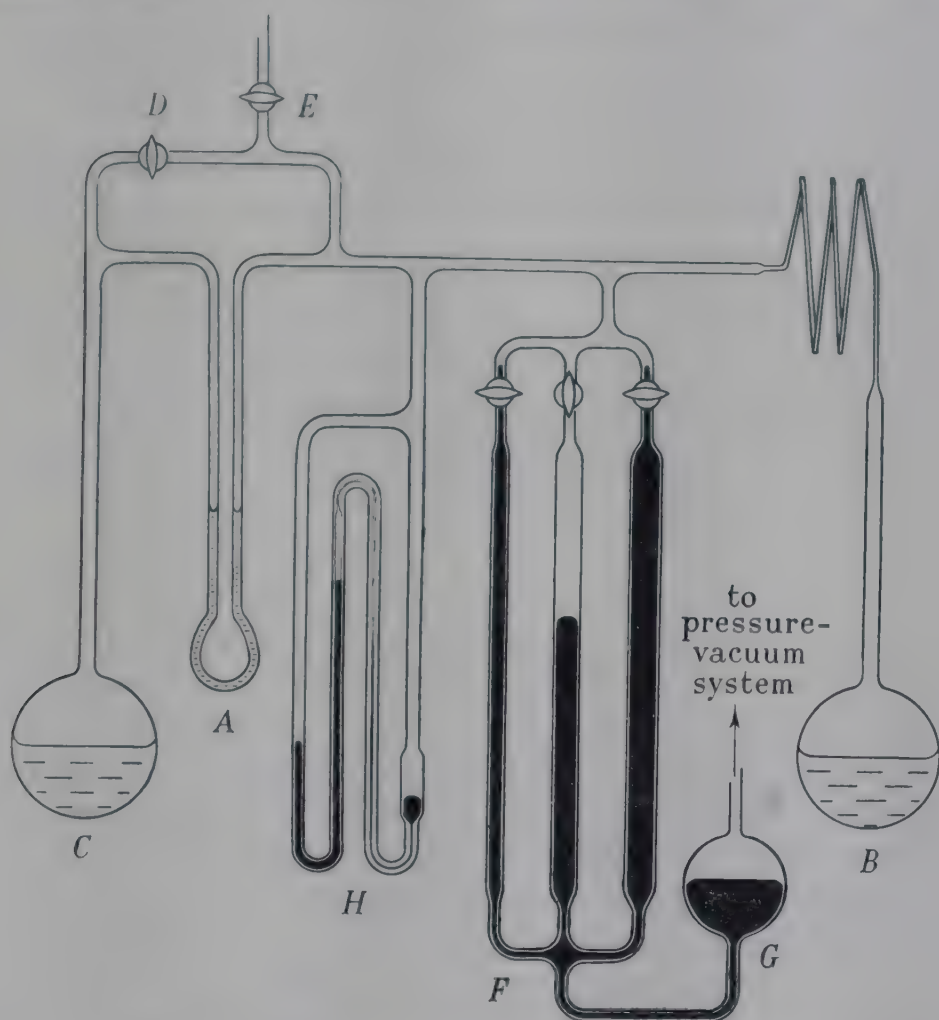


FIGURE 1. Constant pressure absorption apparatus.

A mercury manometer *H* (see Zimmerli 1938) measured pressures in the apparatus below 200 mm. Higher pressures were calculated to the nearest mm. from the mercury levels in *F* and *G* and the air pressure over *G*. Above 200 mm., the oxidation rate was found to be comparatively insensitive to pressure change, and this latter method was then sufficiently accurate.

The apparatus was thus of glass throughout, and being mounted on a single retort stand could be prepared for an experiment on the bench and then completely immersed in the thermostat. The thermostat was maintained at the required temperature to within $\pm 0.02^\circ\text{C}$. The majority of the experiments were carried out at 25°C where the rates of oxidation were most convenient for observation. To obtain the activation energies of the reactions taking place, a second series of experiments was performed at 35°C . This was the highest practical working temperature

at which the initial oxidation rates, now increased by a factor of 4 to 5, could be determined. Though the temperature range was limited, the activation energies determined are undoubtedly of the right order.

(ii) *Materials*

The glucose was of A.R. quality and had a melting-point of 147°C. Sodium hydroxide was also of A.R. quality and did not contain detectable amounts of metals. Solutions were made up and maintained CO₂-free. Sodium bromide was a pure commercial product which was tested and found free from other halides, water and foreign metals. All the solutions were made up with water twice distilled, the last distillation being from a Pyrex still containing KMnO₄.

Oxygen was condensed from gas cylinders, after freezing out H₂O and CO₂, and fractionated. The middle third only was collected for use.

(iii) *Procedure*

In each experiment the compensation and reaction bulbs contained 10 ml. solution. The glucose was present only in the reaction vessel. Throughout the filling operations, the bulbs were frozen in a solid CO₂-ether mixture to prevent reaction taking place. The filled bulbs were sealed on to the apparatus which was then completely evacuated. Oxygen was admitted and the apparatus placed in the thermostat.

After 10 min. standing to allow the main temperature changes to be completed the reaction bulb was shaken and the readings started. Within 30 min. temperature and pressure equilibria were established; the zero reading was then determined by extrapolation to zero time. This procedure was employed because of its comparative simplicity, and it was justified as follows. The course of the oxidation in its initial stages was observed in a few experiments in which it was possible to introduce the glucose within the closed system after equilibrium was reached. For this purpose the glucose solution (containing sodium bromide to render it isotonic with the alkali) was placed in a small tube held upright in the reaction bulb by a thin glass spike extending into the bulb stem. When the bulb was shaken, the glass spike was broken and the solutions mixed. It was then found that the oxidation started immediately and followed a first-order oxygen uptake curve. The extrapolation made in the standard method was shown to estimate the volume of oxygen absorbed in these initial stages quite accurately.

Subsidiary experiments showed that the rate of shaking was not a controlling factor in the rate of oxidation.

Results

(i) *Stoichiometry of the oxidation*

Spengler & Pfannenstiel and Isbell assumed that the stoichiometric course of the autoxidation was represented by the following equation:



To confirm this, a number of oxidations were continued for a prolonged period at atmospheric pressure until the absorption of oxygen had practically ceased. The

glucose concentration was 0.01 mole l^{-1} and the NaOH concentration varied between 0.10 and 1.04 mole l^{-1} . It was found that about 1.05 moles of oxygen were taken up per mole of glucose. Electrometric titrations of the fully oxidized solutions showed that 2.0 moles of NaOH had been neutralized per mole of glucose.

To investigate the oxidation products more closely, a number of oxidations were carried out on a larger scale both with glucose and fructose. A solution of the hexose containing 0.05 mole in 200 ml. of 0.85 N-KOH was shaken vigorously with oxygen under atmospheric pressure in a 500 ml. flask for 2 days. The solution was then treated with an equivalent amount of hydrochloric acid and volatile acids separated by vacuum distillation. The distillates were shown to contain formic acid together with a little lactic acid. These were estimated by titration with alkali and alkaline permanganate. The residue, which consisted mainly of arabonic acid, was dissolved in water and titrated electrometrically. The results are given in table 1.

TABLE 1. ACID PRODUCTS OF THE AUTOXIDATION OF GLUCOSE AND FRUCTOSE

$p\text{O}_2 = 760 \text{ mm.}; 25^\circ \text{ C}$				
moles acid/mole hexose				
hexose	residue	total	distillate	
			formic acid	lactic acid
glucose	1.03	1.00	0.97	0.03
fructose	1.25	1.06	1.03	0.03

It has thus been confirmed that formic acid is a main product, and we conclude that equation (1) is a close approximation to the stoichiometric course of the reaction. These results also show that degradation of glucose into lactic and saccharinic acids is practically eliminated by the presence of oxygen at atmospheric pressure. The slightly high oxygen consumption was probably due to secondary oxidations, and rate determinations were therefore confined to the early part of the oxidation.

(ii) *Dependence of rate of oxidation on glucose concentration*

The absorption of oxygen at pressures of 500 mm. and above followed a uni-molecular course, the rate being proportional to the concentration of glucose in the solution $[S]$ calculated from equation (1). This is demonstrated in figure 2, where $\log [S]$ is plotted against time. At lower pressures of oxygen the first-order coefficient slowly increased with time (see figure 2). This was undoubtedly due to the partial conversion of glucose to fructose which, as is shown later (see (vii)), oxidizes more readily than glucose. These results demonstrate that interconversion, as well as degradation, is suppressed at high oxygen pressures.

The first-order dependence on glucose concentration at high and low pressures was confirmed by measuring the initial rates of oxidation in solutions containing different amounts of glucose. The rate of oxygen absorption per mole of glucose was almost constant but showed a definite downward trend with increase in glucose concentration (table 2). The deviation from a strictly first-order relationship appears

to be connected with the formation of peroxides which enter into secondary reactions. This point is discussed further in a later section.

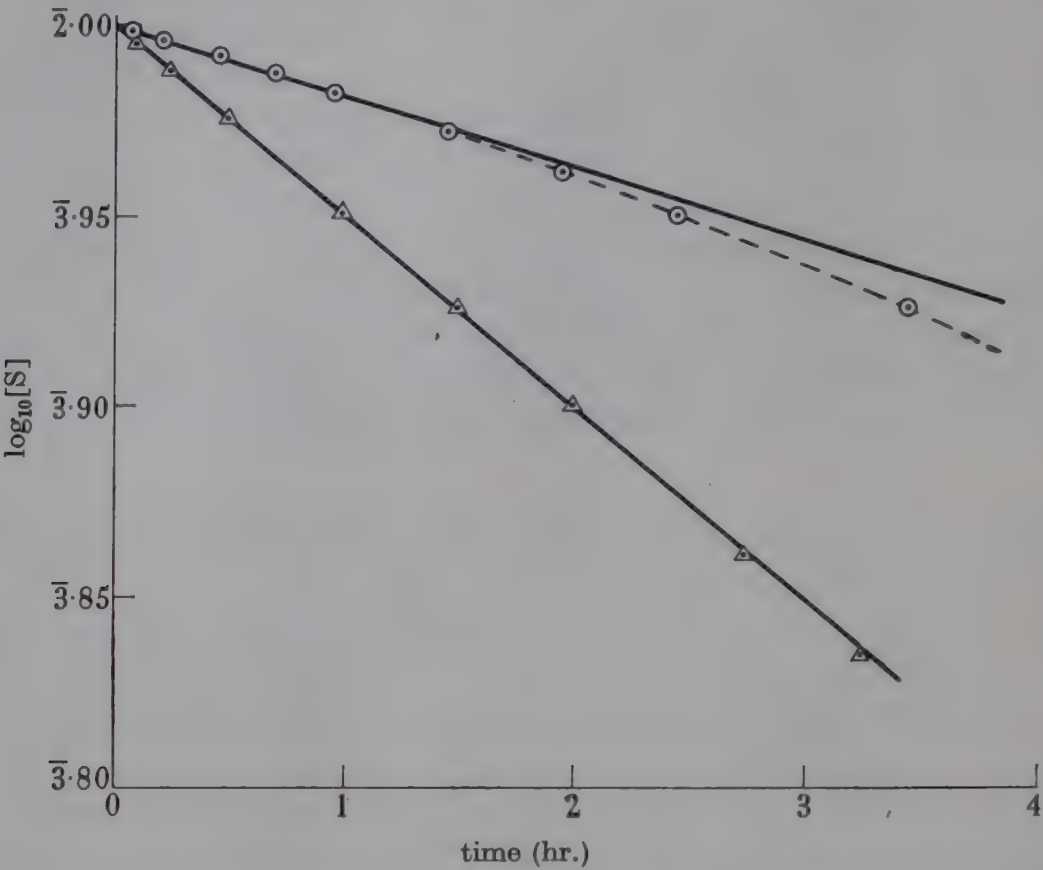


FIGURE 2. Plot of $\log_{10}[S]$ against t for $[S_G]=0.01$ mole $l.^{-1}$, $[NaOH]=1.04$ mole $l.^{-1}$, $25^\circ C$.
 $\odot$ $pO_2=30$ mm.; $\triangle$ $pO_2=460$ mm.

TABLE 2. DEPENDENCE OF RATE OF OXYGEN ABSORPTION ON GLUCOSE CONCENTRATION $25^\circ C$

NaOH (mole $l.^{-1}$)	glucose (mole $l.^{-1}$)	rate of O_2 absorption (mole $l.^{-1}sec.^{-1}$) $\times 10^5$ per mole glucose	
		$pO_2=500$ mm.	$pO_2=55$ mm.
1.04	0.001	4	—
	0.005	3.52	—
	0.010	3.58	1.46
	0.020	3.26	1.39
	0.030	3.09	1.35
	0.040	3.02	—
6.22	0.010	3.79	—
	0.020	3.30	—
	0.030	3.10	—
	0.040	2.96	—

(iii) *Dependence of rate on oxygen pressure and alkali concentration*

The oxygen pressure was varied within the range 25 to 800 mm., and the rate of oxidation was found to fit accurately an expression of the form

$$\text{rate} = \frac{C[S_G]pO_2}{C' + pO_2}, \tag{2}$$

in which C is dependent only on the hydroxyl concentration and temperature, C' is nearly independent of alkali concentration, and $[S_G]$ is the initial sugar concentration. The oxygen dependence is illustrated by the typical reciprocal plot shown in figure 3. The value of C' at 25°C increases from 71 mm. in 0.05 N-NaOH to 79 mm. in 6.22 N-NaOH. At 35°C the corresponding values are 112 and 96 mm. The values of C in the range of alkali concentration studied are given in table 3.

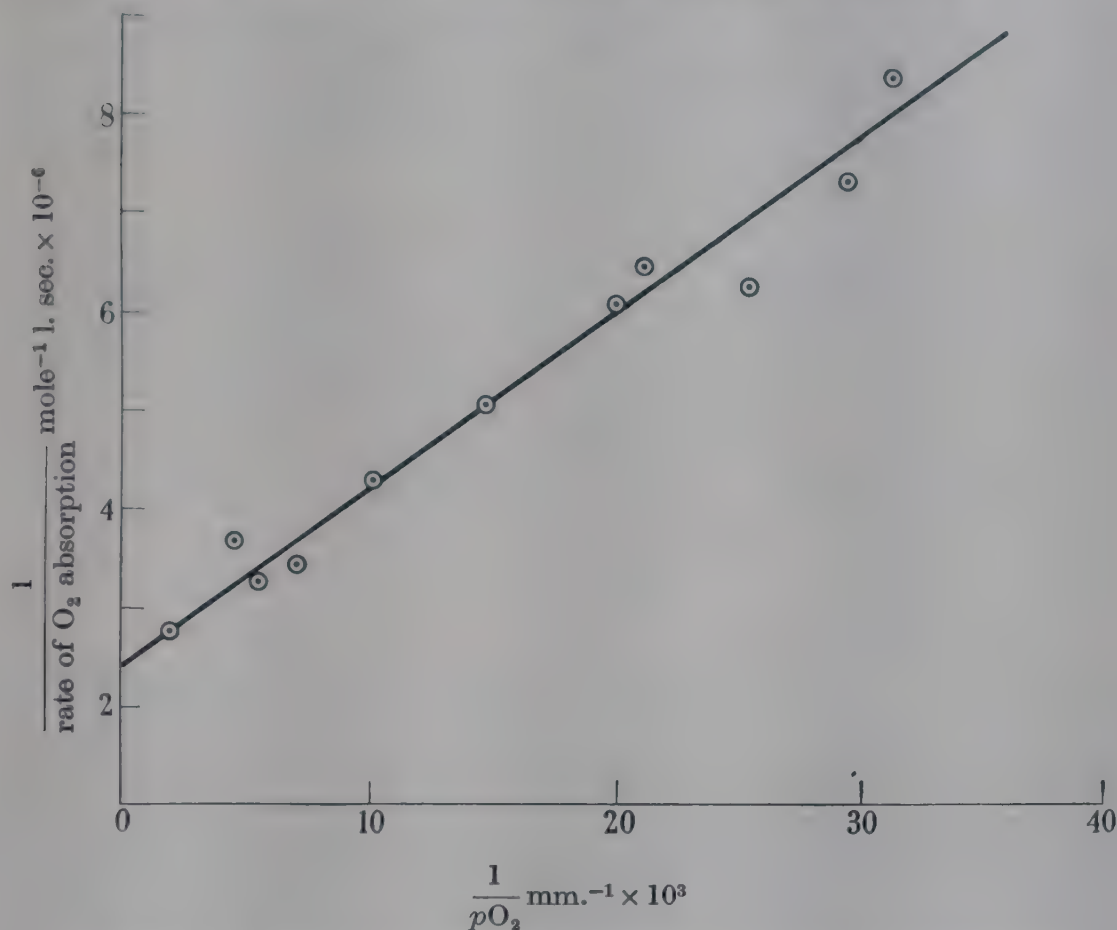


FIGURE 3. Plot of reciprocal rate against reciprocal oxygen pressure for $[S_G] = 0.01$ mole l.⁻¹, $[\text{NaOH}] = 1.04$ mole l.⁻¹, 25°C .

TABLE 3. LIMITING RATES OF OXYGEN ABSORPTION (C)

$[S_G] = 0.01$ mole l.⁻¹; 25°C .

NaOH (mole l. ⁻¹)	C sec. ⁻¹ $\times 10^5$	
	25°C	35°C
0.05	1.67	7.0
0.10	2.42	—
0.15	2.64	—
0.20	2.94	15.6
0.30	3.05	—
0.40	3.44	—
0.50	3.86	18.8
0.75	4.10	19.8
1.04	4.22	21.7
2.00	4.55	—
4.00	4.37	—
6.22	4.41	—

(iv) *Search for possible catalysts and inhibitors*

The addition of benzene diazonium hydroxide or benzoyl peroxide in concentrations of 10^{-3} to 10^{-2} mole l.⁻¹ had no perceptible effect on the rate of oxidation in 1.04 N-NaOH at an oxygen pressure of 500 mm., suggesting that under these conditions the reaction is not readily sensitized by free radicals. The addition of inhibitors such as sulphur, hydroquinone and picric acid was without effect.

Silver, cuprous, ferrous and manganous sulphates in concentrations of 10^{-5} to 10^{-2} mole l.⁻¹ had no obvious catalytic effect under similar conditions; this was also true of cerous nitrate (2.5×10^{-4} mole l.⁻¹) in 6.22 N-NaOH. However, small but definite effects were observed with cobaltous sulphate and platinum black and are summarized in table 4.

TABLE 4. PERCENTAGE INCREASE IN RATE OF O₂ ABSORPTION PRODUCED BY ADDITION OF CoSO₄ (2.5×10^{-4} mole l.⁻¹) AND PT BLACK AT 25° C

catalyst	NaOH (mole l. ⁻¹)	glucose (mole l. ⁻¹)	
		0.01	0.04
CoSO ₄	1.04	+ 15 %	+ 57 %
(pO ₂ = 500 mm.)	6.22	+ 30 %	+ 91 %
CoSO ₄	1.04	+ 30 %	—
(pO ₂ = 85 mm.)			
Pt black	6.22	- 8 %	- 30 %
(pO ₂ = 500 mm.)			

The presence of Co⁺⁺ and Pt black increased the total oxygen absorption per mole of glucose; this tended towards 2 moles with Co⁺⁺ and towards 1.5 moles with Pt black. Thus these substances alter the course of the reaction. Since peroxides are formed during the course of uncatalyzed oxidations, it is conceivable that in the presence of Co⁺⁺ or Pt black, free radicals are generated and may in part be responsible for the effects observed. However, we do not wish to stress this point as there are alternative explanations (cf. the autoxidation of ascorbic acid catalyzed by copper, Silverblatt, Robinson & King 1943) and the elucidation of the mechanism of the catalyzed reaction would require much further work. As would be expected, negligible amounts of peroxide were formed during the course of oxidations with these substances present initially, but when added within the apparatus by means of a magnetic tip during the course of an oxidation, different effects were observed. Upon the addition of CoSO₄, oxygen was not evolved but the rate of oxygen absorption was immediately increased, the increase being sustained. On the other hand, addition of Pt black caused an immediate evolution of oxygen approximately equal to half the peroxidic oxygen content of the solution, followed by a lower rate of oxygen absorption (see figure 4).

There does not seem to be any reason for believing that the uncatalyzed oxidation is a radical chain process, since the only substances which do change the rate also alter the course of the oxidation. Further, if the process were a chain reaction of normal type with long chains, its rate would be given for each glucose species by an expression of the form

$$\sqrt{I} \frac{C_1[G][O_2]}{C_2[G] + [O_2]},$$

where C_1 , C_2 are combinations of velocity constants, I is the rate of chain initiation, and $[G]$ refers to the glucose species under consideration. This type of expression is unsatisfactory for several reasons. (1) The first-order dependence of rate on sugar concentration would not be expected unless $C_2[G]$ were always small compared to $[O_2]$, and I did not involve $[G]$. The former condition does not hold at low oxygen

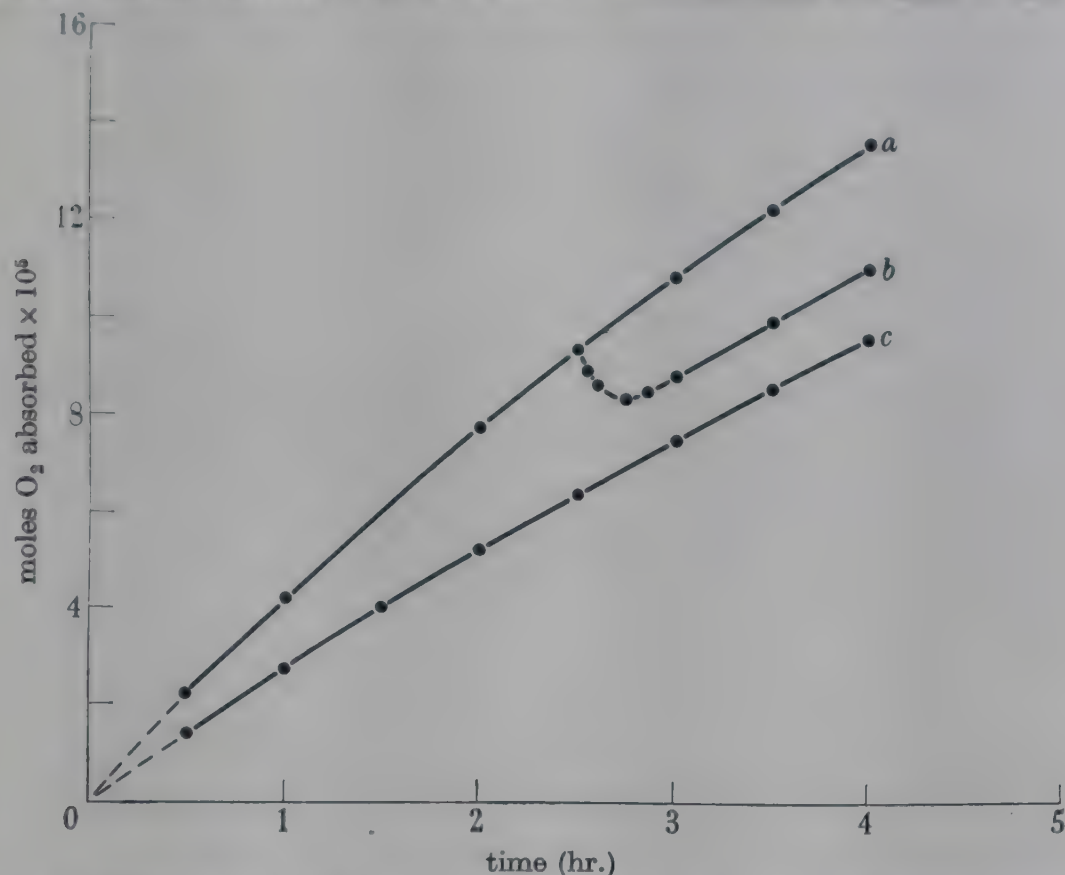


FIGURE 4. Effect of Pt black on the course of the oxidation $[NaOH] = 6.22$ mole l^{-1} . $[S_G] = 0.04$ mole l^{-1} , $pO_2 = 500$ mm. $25^\circ C$. *a* normal oxidation, *b* Pt black added at 2.5 hr., *c* Pt black added initially.

pressures, while the latter does not seem reasonable. (2) The rate of initiation would similarly have to be independent of oxygen pressure. (3) Since the relative amounts of the various glucose species is a function of OH^- , the latter would appear in the denominator of the rate expression. This is not consistent with observation (see equation (2)). The rate of initiation could only be independent of glucose and oxygen concentrations if a surface process were occurring. However, increasing the surface/volume ratio by a factor of 5 to 1000 by the addition of glass beads, powdered glass or $BaSO_4$ (*blanc fixe*) produced no effect on the rate. Thus it seems most likely that the reaction involves ionic intermediates.

(v) Peroxide formation

The presence of peroxide in the partly oxidized solutions was easily demonstrated by the starch-iodide test, as found by Shaffer & Harned (1931). Attempts to isolate the peroxide by solvent extraction or vacuum distillation were not successful, but the concentration was determined in a very reproducible manner by the starch-iodide method in strong acetic acid solution (see Skellon & Wills 1948). In this determination it was assumed that 1 mole of the peroxide oxidized 2 moles of iodide.

At high concentrations of alkali (1 N and above) the initial rate of formation of peroxide is equivalent to the rate of oxygen absorption. At lower alkali concentrations (e.g. 0.10 N) this is no longer true, and the initial rate of peroxide formation is considerably lower than that of oxygen uptake. This is illustrated in table 5.

TABLE 5

$[S_G] = 0.04 \text{ mole l.}^{-1}$; $pO_2 = 500 \text{ mm.}$; 25° C.

NaOH (mole l. ⁻¹)	initial rates (mole l. ⁻¹ sec. ⁻¹) $\times 10^6$	
	peroxide formation	oxygen absorption
0.10	0.10	0.84
1.04	1.0	1.21
6.22	1.1	1.18

As the peroxide accumulates in the system its decomposition becomes increasingly important, with the result that after the initial stages of the reaction have been passed the rate of increase in peroxide concentration falls progressively below the rate of oxygen absorption. The peroxide concentration in the later stages of the reaction are shown in table 6, for a variety of alkali and glucose concentrations, and oxygen pressures.

TABLE 6

[NaOH] (mole l. ⁻¹)	[S _G] (mole l. ⁻¹)	$pO_2 = 480 \text{ mm.}$		$pO_2 = 32 \text{ mm.}$	
		obs.	calc.	obs.	calc.
0.10	0.01	0.38 (5)	0.23	—	—
0.10	0.04	0.67	0.66	—	—
0.15	0.01	0.51	0.33	—	—
0.20	0.01	0.54	0.46	—	—
0.30	0.01	0.67	0.68	0.67 (6.5)	0.41
0.40	0.01	0.95	0.88	0.77 (6)	0.51
0.50	0.01	1.00	1.05	0.83 (6)	0.63
0.75	0.01	1.38	1.41	0.93 (5)	0.77
1.04	0.01	1.67	1.66	1.14 (4.5)	0.90
1.04	0.04	3.42	5.70	—	—
2.00	0.01	1.92	2.04	1.30	1.15
4.00	0.01	2.44	2.32	1.68	1.41
6.22	0.01	2.54	2.53	1.73	1.53
6.22	0.04	6.70	8.00	—	—

The peroxide was estimated at $t = 4 \text{ hr.}$, except in those cases where the time in hours is indicated by a figure in brackets.

The calculated values of the peroxide concentrations given in table 6 will be referred to later.

(vi) Solubility of oxygen

The solubility of oxygen was determined in sodium hydroxide and sodium bromide solutions similar to those used in the oxidation experiments to allow the calculation of velocity constants in absolute units. The results, obtained by degassing oxygen-saturated solutions in a Van Slyke apparatus, are shown in table 7.

TABLE 7. ABSORPTION COEFFICIENTS OF OXYGEN IN SODIUM HYDROXIDE AND SODIUM BROMIDE SOLUTIONS

solution (mole l. ⁻¹)		Bunsen absorption coefficients	
NaOH	NaBr	25° C	35° C
0.00	6.22	0.00251	0.00229
1.04	5.18	0.00229	0.00202
6.22	0.00	0.00177	0.00159

(vii) *Fructose and α -methyl glucoside*

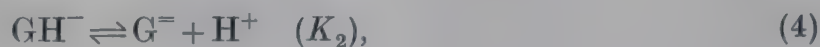
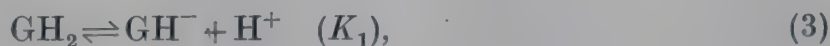
To confirm specific points in the proposed reaction scheme it was necessary to know the rates of autoxidation of fructose and methyl glucoside.

At an oxygen pressure of 500 mm., a solution of fructose (0.01 mole l.⁻¹) in 1.04 N-NaOH absorbs oxygen at about five times the rate at which oxygen is absorbed by a glucose solution under similar conditions at 25° C. It would therefore be expected that the rate of oxidation of a solution of glucose would increase when interconversion to fructose takes place simultaneously. This was found to be the case when low oxygen pressures are used (see (ii)).

Under the same conditions, a solution of α -methyl glucoside (0.04 mole l.⁻¹) showed no absorption of oxygen in 24 hr. Since this compound does not mutarotate, it may be deduced that ionization and possibly ring opening at the terminal carbon atom is necessary for autoxidation to occur. This point is discussed further in part II.

Discussion

Under the conditions of the experiments, glucose is appreciably ionized, and at a given pH there are three or more species in the solution, the concentrations of which are proportional to the total glucose concentration (see Stearn 1931; Urban & Shaffer 1932; and Urban & Williams 1933). These will be represented by GH_2 , GH^- and G^- , being unionized, singly ionized and doubly ionized glucose respectively. Their equilibria are given by



and the following relationships may easily be derived:

$$[\text{GH}_2] = \frac{[\text{S}_G]}{1 + \alpha_1[\text{OH}^-] + \alpha_1\alpha_2[\text{OH}^-]^2}, \quad (5)$$

$$[\text{GH}^-] = \alpha_1[\text{GH}_2][\text{OH}^-], \quad (6)$$

$$[\text{G}^-] = \alpha_1\alpha_2[\text{GH}_2][\text{OH}^-]^2, \quad (7)$$

$$[\text{OH}^-] = c - [\text{GH}^-] - 2[\text{G}^-], \quad (8)$$

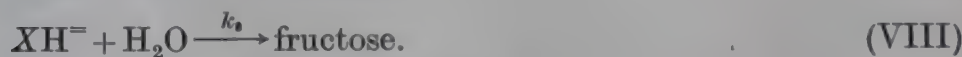
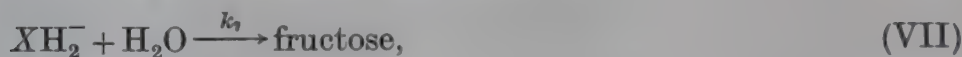
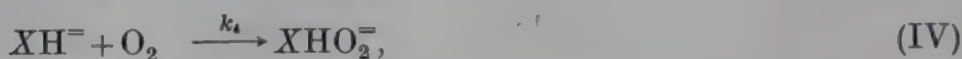
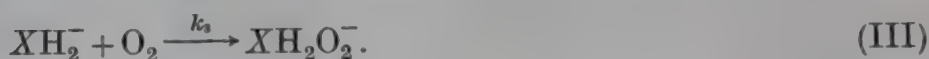
where

$$\frac{K_1 f_1 f_4}{K_w f_2} = \alpha_1, \quad \frac{K_2 f_2 f_4}{K_w f_3} = \alpha_2,$$

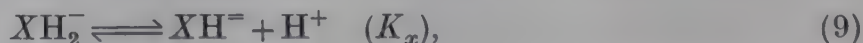
and f_1, f_2, f_3, f_4 are the respective activity coefficients of GH_2 , GH^- , G^- and OH^- . The initial concentration of NaOH is denoted by c . With the small concentrations of

glucose used, $[\text{OH}^-]$ does not differ greatly from c except at the lowest values of c . The values of K_1 and K_2 are not known with any great accuracy but reasonable values are $K_1/K_w = 80 \text{ l.mole}^{-1}$ ($\Delta Q = 4.3 \text{ kcal.}$, see Smith 1936) and $K_2/K_w = 1.5 \text{ l.mole}^{-1}$ ($\Delta Q = 0$ assumed) at 25°C . These values have been adopted for α_1 and α_2 respectively, since it seems reasonable to assume that the combinations of activity coefficients do not differ greatly from unity. In addition, it may be assumed that the activity coefficients are constant throughout the range of $\text{NaOH} + \text{NaBr}$ mixtures of constant ionic strength. The general interpretation of the results is not altered significantly by changing the values of α_1 and α_2 by 50 %.

It is obvious that none of the species GH_2 , GH^- , G^- react directly with oxygen, since the rate is not proportional to the oxygen concentration but to $p\text{O}_2/(\text{constant} + p\text{O}_2)$. The reactive species are thus formed in rate-determining steps at high oxygen pressures. In the following scheme the reactive intermediates XH_2^- and XH^- are formed by the reversible reactions I, II involving GH_2 , GH^- and hydroxyl ions. Kinetically, reversible reactions between GH^- , G^- and water would give identical results, but as appears later (part II), equations (I) and (II) are preferable.



It may be assumed that the ions XH_2^- , XH^- differ only by a proton derived from a hydroxyl group. In this case equilibrium between the ions will be established instantaneously, i.e.



whence
$$[\text{XH}^-] = \frac{K_x}{K_w} [\text{XH}_2^-] [\text{OH}^-] \frac{f_{\text{XH}_2^-} f_4}{f_{\text{XH}^-}} \equiv \alpha_3 [\text{XH}_2^-] [\text{OH}^-]. \quad (10)$$

The autoxidation reactions are represented by equations (III) and (IV) and lead to the formation of peroxide ions which decompose in a series of reactions which are, in general, not rate determining, into arabonate and formate ions (see, however, p. 78). Since at high oxygen pressures the formation of fructose and of lactic and saccharinic acids does not occur appreciably, the latter reactions probably occur through the same intermediates XH_2^- and XH^- (see equations (V) to (VIII)). The kinetics of the autoxidation can be explained without these reactions, since they only appear in the sums $k'_1 + k_5 + k_7$ and $k'_2 + k_6 + k_8$. [However, they are included for the sake of completeness and in anticipation of the results of part II.

The measured rate of oxidation in the steady state is the rate of solution of oxygen. The rate of solution of oxygen is proportional to the oxygen deficiency of the solution; thus, if s is the proportionality constant, $[O_2]_M$ the solubility of oxygen in the given solution and $[O_2]$ the concentration in the liquid during the oxidation, then

$$s([O_2]_M - [O_2]) = k_3[XH_2^-][O_2] + k_4[XH^-][O_2]. \quad (11)$$

Where the rate of shaking is high, s is large and $[O_2]$ does not differ significantly from $[O_2]_M$.

From equations (I) to (VIII) we have

$$\begin{aligned} d([XH_2^-] + [XH^-])/dt = & k_1[GH_2][OH^-] + k_2[GH^-][OH^-] - (k'_1 + k_5 + k_7)[XH_2^-][H_2O] \\ & - (k'_2 + k_6 + k_8)[XH^-][H_2O] - k_3[XH_2^-][O_2] - k_4[XH^-][O_2]. \end{aligned} \quad (12)$$

Application of the stationary state method to equation (12) and insertion of the values of $[GH^-]$ and $[XH^-]$ from equations (6) and (10) gives

$$[XH_2^-] = \frac{[GH_2][OH^-](k_1 + \alpha_1 k_2[OH^-])}{k_3(r_1 + [O_2]) + \alpha_3 k_4[OH^-](r_2 + [O_2])}, \quad (13)$$

where
$$r_1 = \frac{(k'_1 + k_5 + k_7)[H_2O]}{k_3}, \quad r_2 = \frac{(k'_2 + k_6 + k_8)[H_2O]}{k_4}. \quad (14)$$

Thus equation (11) for the rate of oxidation becomes

$$s([O_2]_M - [O_2]) = \frac{[GH_2][OH^-][O_2](k_1 + \alpha_1 k_2[OH^-])(k_3 + \alpha_3 k_4[OH^-])}{k_3(r_1 + [O_2]) + \alpha_3 k_4[OH^-](r_2 + [O_2])}. \quad (15)$$

Alternatively, if XH_2^- and XH^- have different structures so that the equilibrium (9) does not exist we may derive the following three equations:

$$[XH_2^-] = \frac{k_1[GH_2][OH^-]}{k_3(r_1 + [O_2])}, \quad (16)$$

$$[XH^-] = \frac{\alpha_1 k_2[GH_2][OH^-]^2}{k_4(r_2 + [O_2])}, \quad (17)$$

$$s([O_2]_M - [O_2]) = [GH_2][OH^-][O_2] \left\{ \frac{k_1}{r_1 + [O_2]} + \frac{\alpha_1 k_2[OH^-]}{r_2 + [O_2]} \right\}. \quad (18)$$

Both equations (15) and (18) fit the observed first-order dependence of the rate upon glucose concentration, but the oxygen dependence appears as the sum of two terms, each being of the required form (see the empirical equation (2)). If $r_1 \neq r_2$, a reciprocal rate-pressure plot, as in figure 3, would not be linear at values of $[OH^-]$ where both terms are significant. In fact, such curves are linear within experimental error at all values of $[OH^-]$ studied. With $r_1 \approx r_2$, the theoretical equations are in accordance with the observations.

If $r_1 = r_2 = r$, equations (15) and (18) reduce to the same form

$$s([O_2]_M - [O_2]) = \frac{[GH_2][OH^-][O_2](k_1 + \alpha_1 k_2[OH^-])}{r + [O_2]} = -\frac{d[S_G]}{dt}. \quad (19)$$

Equation (2) is identical with (19) if

$$C = (k_1 + \alpha_1 k_2 [\text{OH}^-]) [\text{GH}_2] [\text{OH}^-] / [\text{S}_\text{G}] \quad (20)$$

According to this latter equation, $C[\text{S}_\text{G}]/[\text{GH}_2][\text{OH}^-]$ is linear in $[\text{OH}^-]$, and this is demonstrated in figure 5 for the range of NaOH concentrations of 0.05 to 1.04 mole l.⁻¹. (Equations (5) to (8) are used to evaluate $[\text{GH}_2]$ and $[\text{OH}^-]$.) Extension of the plot to $[\text{NaOH}] = 6.22$ mole l.⁻¹ shows a gradual departure from linearity, the value of $C[\text{S}_\text{G}]/[\text{GH}_2][\text{OH}^-]$ at the highest alkalinity being 21 % below the value obtained by extrapolation of the linear part. This may be due to one or both of the following

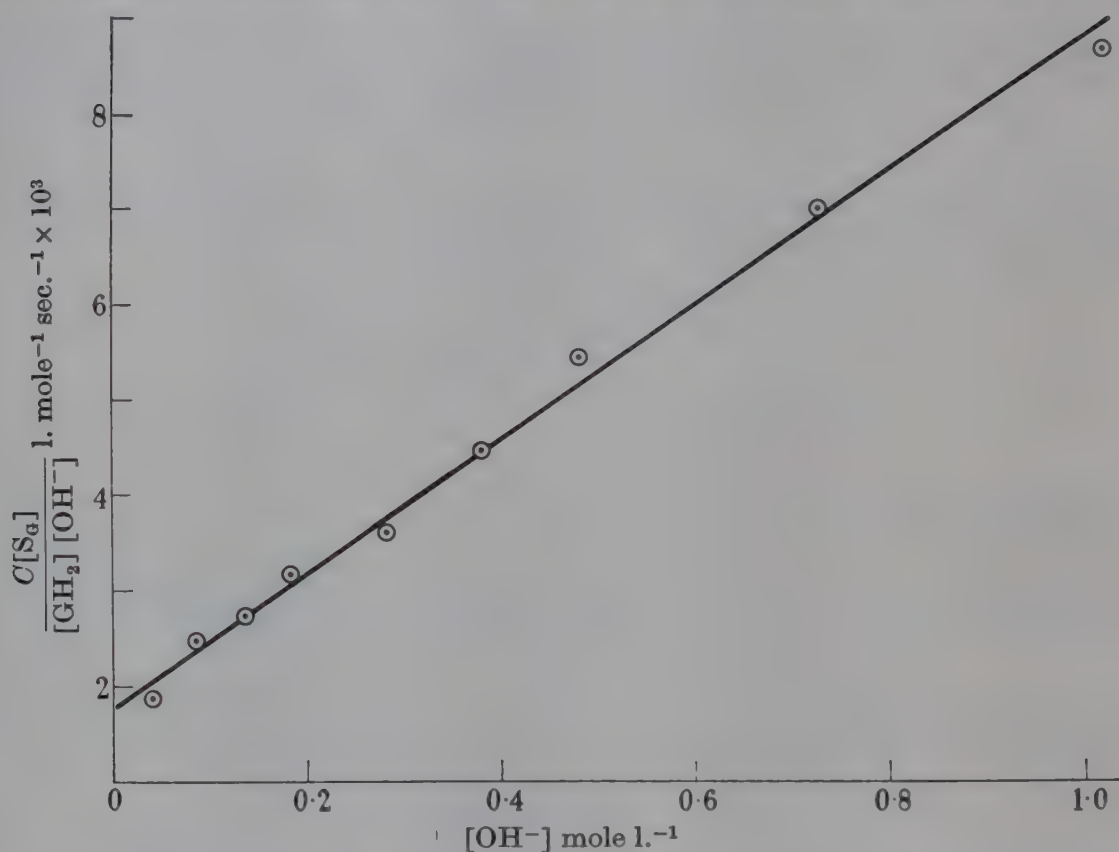


FIGURE 5. Relation of the limiting rate of oxidation to the hydroxyl concentration.

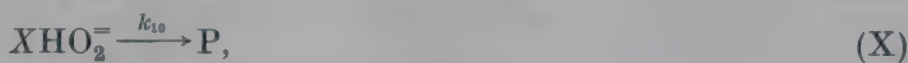
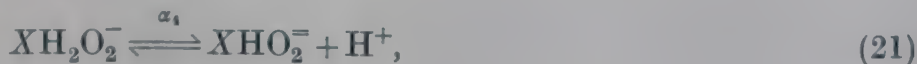
reasons. (i) The extent of the ionization of glucose is not known with certainty. A third ionization constant such that $K_3/K_w = 0.05$ l. mole⁻¹ would indeed produce a linear plot throughout. (ii) There is a considerable change in the nature of the medium between the two ranges of alkali concentration. In the range of 0.05 to 1.04 mole l.⁻¹ hydroxide, the sodium bromide changes from 6.17 to 5.18 mole l.⁻¹ and the medium is thus reasonably constant. This compares with a drop in sodium bromide concentration of from 5.18 mole l.⁻¹ to zero in the hydroxide range of 1.04 to 6.22 mole l.⁻¹. In view of these uncertainties it is felt that equation (19) adequately represents the observed dependence of the rate on the hydroxyl concentration. The kinetic scheme of equations (I) to (VIII) therefore appears to be satisfactorily consistent with the experimental results.

The formation and decomposition of peroxides

The peroxide ions XH_2O_2^- and XHO_2^- formed in reactions (III) and (IV) would be expected to be in equilibrium, since they differ only by a proton derived from

a hydroxyl group. At low concentrations of alkali, the experimental results indicate that only a fraction of the oxygen absorbed appears as a comparatively stable peroxide. It seems likely that the latter is produced from the XHO_2^- ion, since this will be present in relatively small concentrations in weakly alkaline solutions. The singly charged ion must be assumed to decompose rapidly to the final products without forming an intermediate peroxide. The difference in stability between the primary peroxide ions is chemically reasonable. As will appear in part II, both ions are probably derived from open-chain forms of the sugar, but the divalent ion may form the stable peroxide by ring closure.

If we assume that the measured peroxide P decomposes unimolecularly, we obtain the reaction scheme given below:



The stationary state treatment then gives the relation

$$d([\text{XH}_2\text{O}_2^-] + [\text{XHO}_2^-])/dt = R_t - k_9[\text{XH}_2\text{O}_2^-] - k_{10}[\text{XHO}_2^-] = 0, \quad (22)$$

where $R_t = -d[\text{S}_G]/dt$ is the rate of oxygen absorption. Since the variation of $[\text{O}_2]$ is small compared to $[\text{O}_2]$, equation (19) may be integrated giving

$$R_t = R_0 e^{-\rho t}, \quad (23)$$

where R_0 and R_t are the values of $-d[\text{S}_G]/dt$ initially and at time t , and

$$\rho = R_0/[\text{S}_G]_0 = R_t/[\text{S}_G]_t.$$

From equations (21) to (23)

$$[\text{XHO}_2^-] = \frac{R_0 \alpha_4 [\text{OH}^-] e^{-\rho t}}{k_9 + k_{10} \alpha_4 [\text{OH}^-]}. \quad (24)$$

Further,
$$d[\text{P}]/dt = k_{10}[\text{XHO}_2^-] - k_{11}[\text{P}]. \quad (25)$$

Combining equations (24) and (25) and integrating, we obtain for the peroxide concentration $[\text{P}]_t$ at time t

$$[\text{P}]_t = \frac{R_0 [\text{OH}^-] e^{-\rho t} (1 - e^{-(k_{11} - \rho)t})}{(k_{11} - \rho) \{ [\text{OH}^-] + k_9/(k_{10} \alpha_4) \}}. \quad (26)$$

In this equation R_0 and ρ may be evaluated by use of equation (19) and the constants given below in table 8. At alkali concentrations above 1 N and with the higher glucose concentrations, where deviations from the calculated rates of oxidation are significant, the observed rates of oxidation are used to give R_0 and ρ directly.

Values of $[\text{P}]_t$ have been calculated from equation (26) using the following constants:

$$k_9/(k_{10} \alpha_4) = 0.63 \text{ mole l.}^{-1}, \quad k_{11} = 5.5 \times 10^{-5} \text{ sec.}^{-1},$$

and R_0, ρ estimated as above. These values have been given in table 5. The agreement between calculated and observed figures is satisfactory except at the higher glucose

concentration (0.04 M); under these conditions the calculated value exceeds the observed one. It will be recalled that the rate of oxygen uptake in these solutions is rather lower than would be expected on the basis of a first-order dependence on glucose concentration. The two phenomena appear to be interconnected and may be the result of a reaction of the type



between the reducing and oxidizing ions present. A small fraction of the XH_2^- intermediates reacting in this manner would be sufficient to account for the experimental observations.

Values of the velocity constants

From figure 5, k_1 and k_2 may be obtained directly, since k_1 is the intercept and $\alpha_1 k_2$ the slope of the curve. The values of r_1 and r_2 are given, in units of mm. Hg pressure, by the dependence of the oxidation rate upon oxygen pressure as already described, and these may be converted into absolute units by use of the solubility data obtained. The evaluation of the separate velocity constants involved in r_1 , r_2 is discussed in part I, section B and part II; it is obviously not possible from the present results alone. The values of the velocity constants, activation energies and frequency factors are given in table 8. The value of $[\text{H}_2\text{O}]$ has been taken to be 50.0 mole l.⁻¹ the mean value of the concentration in 6.22 N-NaOH and in 6.22 M-NaBr.

TABLE 8

velocity constants	25° C	35° C	<i>E</i> (kcal.)	<i>A</i>
k_1 l. mole ⁻¹ sec. ⁻¹	1.79×10^{-3}	7.39×10^{-3}	25.9	1.6×10^{16}
k_2 l. mole ⁻¹ sec. ⁻¹	8.65×10^{-5}	4.12×10^{-4}	28.5	6.5×10^{16}
$\frac{k'_1 + k_5 + k_7}{k_3}$	2.00×10^{-7}	2.82×10^{-7}	6.3*	—
$\frac{k'_2 + k_6 + k_8}{k_4}$	2.22×10^{-7}	2.42×10^{-7}	1.6*	—

* Apparent.

B. NON-STEADY-STATE EXPERIMENTS

Further separation of the constants is possible if the concentrations of the intermediates can be measured. This may be done by considering non-stationary states.

From equation (13) or alternatively equations (16) and (17), it follows that the stationary concentrations of XH_2^- and XH^- are greater in the absence of oxygen. It may be shown from the results given in table 8 that at zero oxygen pressure, the concentrations of XH_2^- and XH^- are approximately four times as large as under steady state conditions with 225 mm. O_2 . Therefore on the introduction of oxygen to an oxygen-free alkaline glucose solution, the initial rate of oxidation should be greater than the steady rate. This is the effect which was found and measured in the following experiments. It must be emphasized that in calculating absolute velocity constants from these results it is assumed that the excess oxygen absorption is due entirely to reaction with XH_2^- and XH^- , and that any other intermediates which

may take part in the glucose-fructose interconversion do not affect the result. As will appear later the kinetics of the interconversion reactions can be explained in terms of rate determining steps involving XH_2^- , XH^- , and so this assumption is reasonable.

Experimental

The oxygen absorption apparatus and the materials used are as described in section A of this paper. Owing to the small differences in oxygen absorption to be measured (of the order of 0.05 ml.) it was found more convenient to observe the deflexion of the dibutyl phthalate manometer by means of a cathetometer rather than to measure directly the uptake by means of the gas burettes. The deflexion was linear within experimental error and absorptions could be measured to 0.001 ml. with accuracy.

Since the rate of solution of oxygen was a significant quantity in these experiments it was necessary to control both the rate of shaking and the time of shaking as closely as possible. The mechanical shaker was therefore operated by means of a d.c. motor operating off storage batteries which gave a constant speed. The high starting torque of the motor enabled the maximum rate of shaking to be reached within 0.5 sec. and by means of regenerative braking the shaking could be stopped within the same period, which was sufficiently small as to be negligible.

Procedure

The reaction bulb and compensation bulb were filled as described previously. The reaction bulb was shaken as soon as possible after the apparatus had been immersed in the thermostat to reduce the formation of fructose while temperature and pressure equilibria were being established. After 10 min. the shaking was stopped for a pre-determined interval. The oxygen uptake was then measured upon resumption of shaking for a further period followed by a measurement of the steady rate of oxygen absorption over a similar period. The difference between the two absorptions gave the excess amount of oxygen consumed by the solution after a period of standing. During the period of standing the rate of oxygen absorption was reduced to about 2 % of the normal rate; this was sufficiently low to permit the assumption of an oxygen-free solution to be made in the calculations.

It was desirable to keep the rest period as short as possible, consistent with the establishment of equilibrium, to prevent appreciable conversion to fructose. The period of shaking following the rest period was only long enough to allow the new stationary state to be approached closely. This latter was necessary to obtain accurate values of the excess absorption. Suitable values for the rest and shaking periods were estimated from preliminary experiments, and were improved upon as the probable values of the constants involved were revealed.

It was not found possible to obtain more than about six repeats of the 'rest-shake' cycle with one solution before the formation of fructose was demonstrated by an appreciable increase in reaction rate, necessitating a start with a fresh glucose solution. At this point, not more than 5 % of the glucose had been oxidized.

Results

The average excess oxygen absorptions measured are given in table 9.

TABLE 9. EXCESS OXYGEN ABSORPTIONS

$p\text{O}_2 = 180 \text{ mm.}$; frequency of shaking = 450/min.

temp. (° C)	glucose (mole l. ⁻¹)	NaOH (mole l. ⁻¹)	time standing (min.)	time shaking (min.)	excess O ₂ absorption (mole l. ⁻¹ × 10 ⁵)
15	0.25	0.361	6	3	4.2
15	0.25	1.399	6	3	3.5
25	0.05	0.150	3	2	3.1
25	0.05	1.040	3	2	2.5

The excess oxygen uptake is made up of two parts: (a) that which reacts with XH_2^- , XH^- and (b) that required to increase the concentration of oxygen in the solution to its steady-state value at the oxygen pressure employed. The values for the solubility of oxygen in the solution are given in section A.

In order to evaluate the velocity constants of the reactions of XH_2^- and XH^- with oxygen, it is necessary to know the value of s , the rate constant for the solution of oxygen. The rate of increase of oxygen concentration in the liquid is given by equation (28) below:

$$d[\text{O}_2]/dt = s([\text{O}_2]_M - [\text{O}_2]) - k_3[\text{XH}_2^-][\text{O}_2] - k_4[\text{XH}^-][\text{O}_2]. \quad (28)$$

If the oxidation reactions are negligible, equation (28) may be integrated to give equation (29)

$$[\text{O}_2] = [\text{O}_2]_M (1 - e^{-st}). \quad (29)$$

It is not possible to measure s directly by addition of oxygen to the liquid in the absence of glucose, because the rate of solution is so high that the liquid is effectively saturated before pressure equilibrium has been established. However, by using solutions containing a small amount of glucose (0.05 and 0.01 M at 15 and 25°C respectively) it is possible to obtain practically oxygen-free solutions by interrupting the shaking at any time, and so the disturbing effects due to diffusion could be eliminated. The volumes of oxygen absorbed were measured as a function of time which had elapsed since shaking was started.

It was found that the initial portion of the absorption-time curve closely fitted the form of equation (29), and the value for s was found to be 0.20 sec.⁻¹ at 15 and 25°C.

Calculation of the velocity constants

The measured absorption of oxygen by the solution over a period of time t_1 to t_2 is given by

$$A = \int_{t_1}^{t_2} s([\text{O}_2]_M - [\text{O}_2]) dt. \quad (30)$$

The alternative reaction schemes mentioned in section A may now be considered separately.

Scheme I. Exclusion of the equilibrium $XH_2 \rightleftharpoons XH^- + H^+$.

From equations (I) to (VIII) and (14) the following general equations may be derived:

$$d[XH_2^-]/dt = k_1[GH_2][OH^-] - k_3(r_1 + [O_2])[XH_2^-], \quad (31)$$

$$d[XH^-]/dt = \alpha_1 k_2[GH_2][OH^-]^2 - k_4(r_2 + [O_2])[XH^-]. \quad (32)$$

Considering these equations with equations (28) and (30), the five unknowns, k_3 , k_4 , $[O_2]$, $[XH_2^-]$ and $[XH^-]$, have to be evaluated from four equations. Since the experiments were carried out at two values of $[OH^-]$ a complete solution is possible. Numerical solutions of equations (28), (30), (31) and (32) were obtained using the differential analyzer. A number of A , t curves were drawn by the analyzer for a series of values of k_3 and k_4 , from $t = 0$, $[O_2] = 0$ (beginning of shaking) until the steady

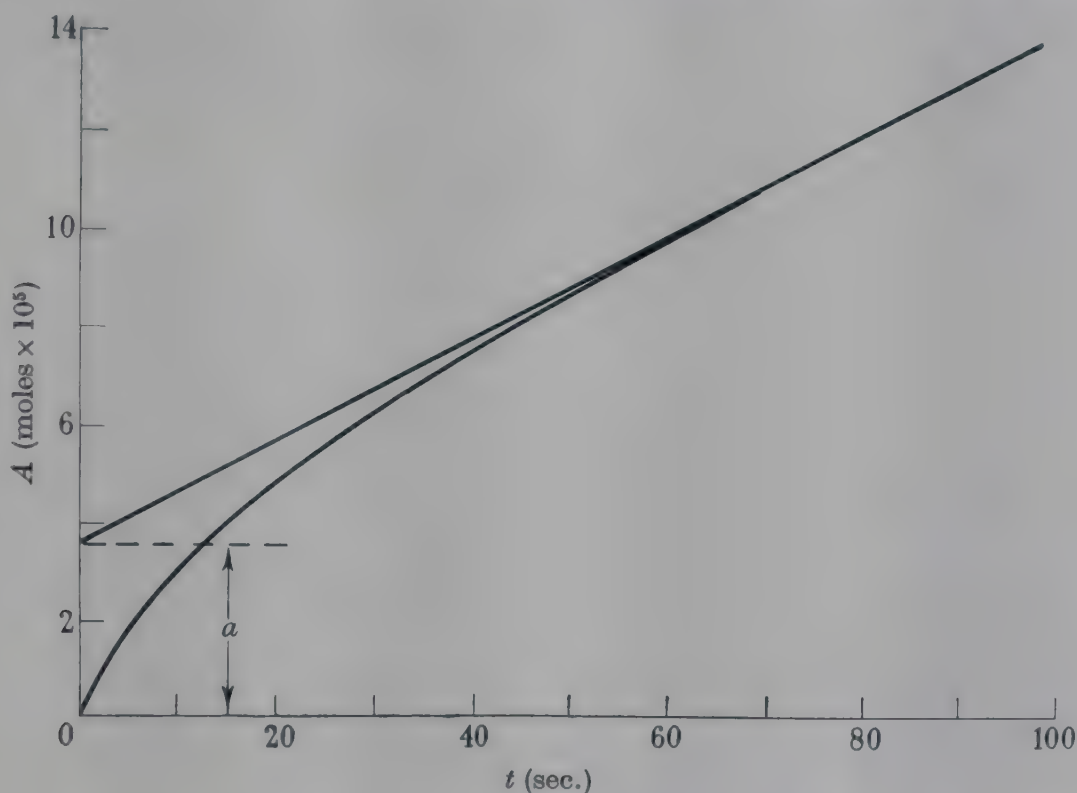


FIGURE 6. Calculated curve of oxygen absorption A against t for $k_3 = 2.0 \times 10^3$ l. mole $^{-1}$ sec. $^{-1}$, $[S_0] = 0.25$ mole l. $^{-1}$, $k_4 = 6.0 \times 10^3$ l. mole $^{-1}$ sec. $^{-1}$, $[OH^-] = 1.00$ mole l. $^{-1}$, $pO_2 = 180$ mm.; 15° C. a = excess oxygen absorption.

absorption rate was reached ($d[O_2]/dt = 0$). Linear extrapolation of the steady rate from the latter point to $t = 0$ gave the excess oxygen absorption which was the experimentally measured quantity. By adjustment of k_3 and k_4 the calculated and experimental values of the excess absorption were brought into agreement (see figure 6).

The values finally adopted for k_3 and k_4 are given in table 10, together with those of $k'_1 + k_5 + k_7$ and $k'_2 + k_6 + k_8$ calculated from table 8. These values are subject to possible errors of $\pm 10\%$ owing to the comparatively small oxygen absorption readings. The activation energies and frequency factors are given as an indication of the order of magnitude.

TABLE 10. VELOCITY CONSTANTS, L. MOLE⁻¹ SEC.⁻¹

	15° C	25° C	<i>E</i> (kcal.)	<i>A</i>
$k'_1 + k_5 + k_7$	2.8×10^{-4}	6.0×10^{-4}	12.9	—
$k'_2 + k_6 + k_8$	1.2×10^{-3}	2.4×10^{-3}	11.8	—
k_3	2.0×10^3	3.0×10^3	6.9	3.5×10^3
k_4	6.0×10^3	10.8×10^3	10.0	2.4×10^{11}

Concentration-time curves for $[O_2]$, $[XH_2^-]$ and $[XH^-]$ were constructed from the analyzer readings and showed that the rest and shaking periods chosen in the experiments were sufficient for effective equilibrium to be reached. Typical curves are shown in figure 7a.

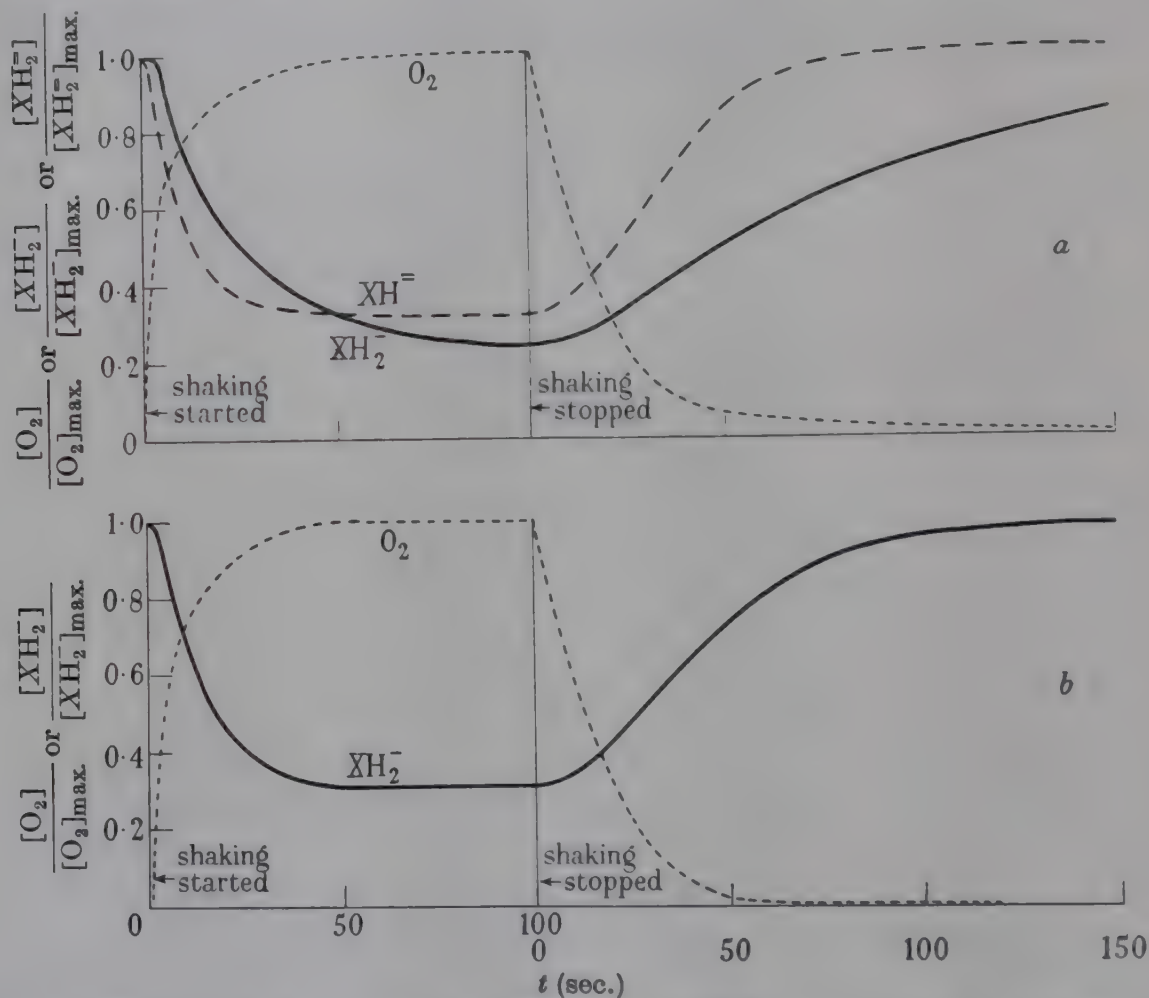


FIGURE 7. Calculated changes in $[O_2]$, $[XH_2^-]$ and $[XH^-]$ in shaking and standing cycle. (a) exclusion of $XH_2^- \rightleftharpoons XH^- + H^+$ equilibrium, (b) inclusion of $XH_2^- \rightleftharpoons XH^- + H^+$ equilibrium, $k_3 = 2.0 \times 10^{-3}$ l. mole⁻¹ sec.⁻¹, $[S_0] = 0.25$ mole l.⁻¹, $k_4 = 6.0 \times 10^{-3}$ l. mole⁻¹ sec.⁻¹, $[OH^-] = 1.00$ mole l.⁻¹; $pO_2 = 180$ mm.; 15° C. $[O_2]/[O_2]_{max.}$ — $[XH_2^-]/[XH_2^-]_{max.}$ $[XH^-]/[XH^-]_{max.}$

Scheme II. Inclusion of the equilibrium $XH_2^- \rightleftharpoons XH^- + H^+$.

From equations (12) and (28) by substitution from equations (10) and (14) we obtain

$$\frac{d[XH_2^-]}{dt} = \frac{1}{1 + \alpha_3[OH^-]} [GH_2][OH^-] (k_1 + \alpha_1 k_2 [OH^-]) - [XH_2^-] (k_3(r_1 + [O_2]) + k_4 \alpha_3 [OH^-] (r_2 + [O_2])), \quad (33)$$

$$d[O_2]/dt = s([O_2]_M - [O_2]) - [XH_2^-][O_2] (k_3 + k_4 \alpha_3 [OH^-]). \quad (34)$$

From equations (30), (33) and (34) it will be seen that there are five unknowns, k_3 , k_4 , α_3 , $[XH_2^-]$ and $[O_2]$. The experiments at two NaOH concentrations give a fourth relationship which is insufficient for a full solution, but the values for k_3 and k_4 determined in scheme I may be used and the effect of varying α_3 determined. The value of α_3 will probably be similar to that of α_2 , referring to the second ionization constant of glucose ($\alpha_2 = 1.5 \text{ l. mole}^{-1}$) and so values of α_3 of 0.1, 1.0 and 10 l. mole^{-1} were tried. Since the differential analyzer was not available for this part of the work, the equations were solved by stepwise integration. The calculated values of excess oxygen absorption were close to the observed values given in table 9, the effect of varying α_3 being negligible. The only difference compared to scheme I is that $[XH_2^-]$ and $[XH^-]$ now approach their equilibrium values at the same rate (see figure 7b), but the difference was not observable experimentally.

There is therefore no means of distinguishing between schemes I and II, the values given in table 10 fitting either equally well.

Comparison with previous results

The present results confirm the observations of Clifton & Ort (1930) and Roepke & Ort (1931) on the rate of reaction of glucose with various oxidants and the concentrations of the intermediates.

Roepke & Ort studied the rate of reaction between potassium ferricyanide and glucose in phosphate buffers by measuring the redox potential. The rate of reaction was independent of the ferricyanide concentration and it was assumed that the measured rate was that of the formation of the 'active reductant'. Over the range of pH 8 to 11 the rate was closely proportional to the OH^- concentration. Examination of their figures for pH 8.01 in M/15 buffer at 30°C shows that the rate of disappearance of ferricyanide per mole of glucose is $4.16 \times 10^{-9} \text{ mole sec.}^{-1}$. Assuming that the molecule of 'active reductant' is completely oxidized by four molecules of ferricyanide we obtain the value $k_1 = 1.04 \times 10^{-3} \text{ l. mole}^{-1} \text{ sec.}^{-1}$ by equation (I). This compares with the interpolated figure of $3.7 \times 10^{-3} \text{ l. mole}^{-1} \text{ sec.}^{-1}$ from the autoxidation experiments. The agreement is good considering the extrapolation and the fact that we are comparing figures obtained on solutions greatly differing in ionic strength. Roepke & Ort found that the oxidation rate increased by a factor of 1.38 by increasing the buffer concentration from M/15 to M/5; in the present experiments the total salt concentration was constant at 6.22.

At pH 10 in a phosphate buffer solution, Clifton & Ort found by potentiometric titration with oxygen that the concentration of the intermediate was one part in 266,000 of glucose at an unspecified temperature, probably 30°C . The assumption was made that the reaction was, in effect, between one molecule of intermediate and one atom of oxygen. This is probably incorrect and their concentration of active reductant should be given as one part in 133,000. Utilizing equation (16) and putting in the values obtained in the present experiments we obtain corresponding values of one part in 168,000 at 25°C and one part in 116,000 at 30°C . This is striking agreement considering the extrapolation involved and the possible experimental errors.

Further, from the comparison with Ort's work, it would appear that the rates of formation and disappearance of XH_2^- are equally accelerated by the addition of salt so that the concentration of the intermediate at equilibrium is nearly independent of the salt concentration.

The mechanism of equations (I) to (VIII) therefore appears to be in complete agreement with all the available experimental observations on the oxidation of glucose. This reaction scheme is in fact one of the simplest which could be devised, having as its essential steps simple prototropic reactions involving the various unionized and ionized species of the sugar which are known to exist in alkaline solution.

The mechanical integration of the equations was carried out under the direction of Mr J. Crank on the differential analyzer built in these laboratories. The authors are also indebted to Dr H. Tompa for mathematical assistance and to Miss M. E. Henry for carrying out the numerical integrations.

REFERENCES

- Bolland, J. L. & Gee, G. 1946 *Trans. Faraday Soc.* **42**, 236.
Clifton, C. E. & Ort, J. M. 1930 *J. Phys. Chem.* **34**, 855.
Dixon, M. 1943 *Manometric methods*. Cambridge University Press.
Evans, W. L. 1929 *Chem. Rev.* **6**, 281.
Evans, W. L. 1942 *Chem. Rev.* **31**, 537.
Glattfeld, J. W. 1913 *Amer. Chem. J.* **50**, 135.
Hammett, L. P. 1940 *Physical organic chemistry*. New York: McGraw-Hill.
Isbell, H. S. 1942 *Bur. Stand. J. Res., Wash.*, **29**, 227.
Isbell, H. S. 1944 *Bur. Stand. J. Res., Wash.*, **32**, 45.
Nef, J. U. 1910 *Liebigs Ann.* **376**, 1.
Nef, J. U. 1914 *Liebigs Ann.* **403**, 204.
Power, M. H. & Upson, F. W. 1926 *J. Amer. Chem. Soc.* **48**, 195.
Roepke, M. H. & Ort, J. M. 1931 *J. Phys. Chem.* **35**, 3596.
Shaffer, P. A. & Friedemann, T. E. 1930 *J. Biol. Chem.* **86**, 345.
Shaffer, P. A. & Harned, B. K. 1931 *J. Biol. Chem.* **93**, 311.
Silverblatt, E., Robinson, A. L. & King, C. G. 1943 *J. Amer. Chem. Soc.* **65**, 137.
Skellon, J. H. & Wills, E. D. 1948 *Analyst*, **73**, 78.
Smith, G. F. 1936 *J. Chem. Soc.* p. 1824.
Spengler, O. & Pfannenstiel, A. 1935 *Z. Ver. deutsch. Zuckerindustr.* **85**, 546.
Stearn, A. E. 1931 *J. Phys. Chem.* **35**, 2226.
Urban, F. & Shaffer, P. A. 1932 *J. Biol. Chem.* **94**, 697.
Urban, F. & Williams, R. D. 1933 *J. Biol. Chem.* **100**, 237.
Wolfson, M. L. & Lewis, W. L. 1928 *J. Amer. Chem. Soc.* **50**, 837.
Zimmerli, A. 1938 *Industr. Engng Chem.* (Anal. ed.), **10**, 283.

Kinetic studies on carbohydrates in alkaline conditions

II. The kinetics of the rearrangements of glucose and fructose in alkaline solution

BY C. H. BAMFORD, DAPHNE BAMFORD AND J. R. COLLINS
Courtaulds Ltd., Maidenhead, Berks

(Communicated by A. H. Wilson, F.R.S.—Received 3 May 1950—
Revised 15 July 1950)

A kinetic study of the degradation and interconversion of glucose and fructose in alkaline solution has been carried out. The change in sugar concentration has been followed by electrometric titration of lactic and saccharinic acids, which are the main products of the degradation in the range of alkali concentration used. A kinetic analysis of the results indicates that all the reactions which occur are of first order with respect to the sugar concentration.

In a general discussion of parts I and II of this series, it is shown that the observed dependence of the rates of the degradation and interconversion reactions of glucose upon the sugar and alkali concentration can be explained in terms of the ionic intermediates postulated for the autoxidation reaction. These intermediates are therefore probably concerned in the rate-determining steps of the reactions in the range of alkali concentration used. The absolute velocity constants of most of the reactions have been calculated on this basis. The structure of the intermediates and the mechanism of the reactions are discussed.

INTRODUCTION

In part I of this series, the kinetics of the autoxidation of glucose have been discussed, and the observations have been shown to be consistent with a mechanism which involves ionic intermediates (XH_2^- , XH^-) formed by prototropic reactions of unionized and ionized glucose. The present paper seeks to demonstrate that the kinetics of the interconversion and degradation reactions of glucose can be accounted for in terms of the same intermediates.

The experiments have been carried out in comparatively concentrated solutions of potassium hydroxide, since under these conditions the degradation reactions leading to lactic and saccharinic acids occur almost quantitatively (compare Shaffer & Friedemann 1930). These acids may be readily estimated and so the course of the degradation may be followed in a simple manner. This is no longer true at low alkali concentrations, since a diversity of products is obtained and accurate estimation is impossible.

EXPERIMENTAL

The potassium hydroxide solutions used in these experiments had concentrations in the range 1 to 5 N. All solutions were made up to a total electrolyte concentration of 5 N with sodium bromide to minimize the salt effect (compare part I).

The reactions were carried out in Pyrex vessels consisting of two compartments, one of which contained initially a solution of the sugar in distilled water and the other potassium hydroxide solution containing sodium bromide. The volumes and con-

centrations of the solutions were arranged so that after mixing the liquid was equivalent to a mixture of 25 ml. 5 N electrolyte and 2 g. sugar, while the concentration of alkali had the desired value. The reaction vessels were sealed after air had been displaced by nitrogen.

Runs were carried out at 25° and 45° C. At 45° C, in view of the short reaction times, the heat developed on mixing the solutions could lead to an appreciable error. The reaction vessel was therefore heated to a temperature (determined by trial) such that, after mixing, the required temperature was attained before immersion in the thermostat. These precautions were unnecessary at 25° C.

At the end of the run, the reaction mixture was rapidly cooled and brought to pH 8 by the addition of 5 N-hydrochloric acid, the liquid being thoroughly cooled during neutralization. The solution was then titrated electrometrically against standard (usually 5 N) HCl using a pH-meter with glass and calomel electrodes, reading to 0.01 of a unit. When the pH had fallen to about 1.5 the liquid was removed and boiled under reflux for 15 min. in an atmosphere of nitrogen. It was then made alkaline, and retitrated with standard HCl. The second titre was always smaller than the first. Subsidiary experiments showed that in runs in which the alkali concentration was greater than 1 N, negligible quantities of CO₂ and volatile acids were produced. The diminution in the acid content on boiling was therefore not due to loss of a volatile acid. Saccharinic acids are major products of the alkaline degradation of sugars, and on boiling under acid conditions these substances lactonize rapidly. It is considered, therefore, that the difference between the two titres represents the saccharinic acid formed in the reaction, while the final titre corresponds to the lactic acid.

The initial OH⁻ concentration could be calculated from the ionization constants of the sugar (see part I), and after complete degradation the final OH⁻ concentration could be estimated from the amounts of lactic and saccharinic acids produced. The differences between the initial and final OH⁻ values were negligible when the alkali concentration was in the range 2 N to 5 N, but the change (0.1 mole l.⁻¹) in the experiments with 1 N alkali was somewhat larger. However, it was not necessary to use conversions greater than 60 % to obtain the relevant data, and while the change in alkali concentration in this case was not negligible, it was probably not very significant.

RESULTS

Experiments were carried out with alkali concentrations of 1, 2, 2.8, 4 and 5 N, and at each concentration lactic and saccharinic acids were estimated for different reaction times. Thus the residual concentration of sugar could be calculated as a function of the time of reaction. The maximum yields of the acids corresponded to at least 94 % conversion of the sugar, except for the reaction in 1 N alkali, where the maximum conversion was only 85 %. In the latter case a correction was therefore applied in calculating the residual sugar concentration. (It has been reported (see e.g. Nef 1910; Upson 1911) that small quantities of 1,3-dihydroxybutyric acid are formed in the degradation under rather more violent conditions than the ones we have used. If this substance were formed in our experiments it would be included

in the saccharinic acid, and the true values for the maximum conversion of the sugars into lactic, saccharinic, and dihydroxybutyric acids would be rather lower than the figures given above. However, it is easily seen that the residual sugar concentration at any stage of the degradation, estimated as described, would not be affected by the presence of dihydroxybutyric acid.)

For the purpose of illustration the results obtained with glucose and fructose in 2 N alkali at 45° C are given below in detail. Figure 1 shows the course of the reaction with time. In each diagram the full curve represents the total (lactic + saccharinic) acids formed, while the broken curve corresponds to the lactic acid alone. From these

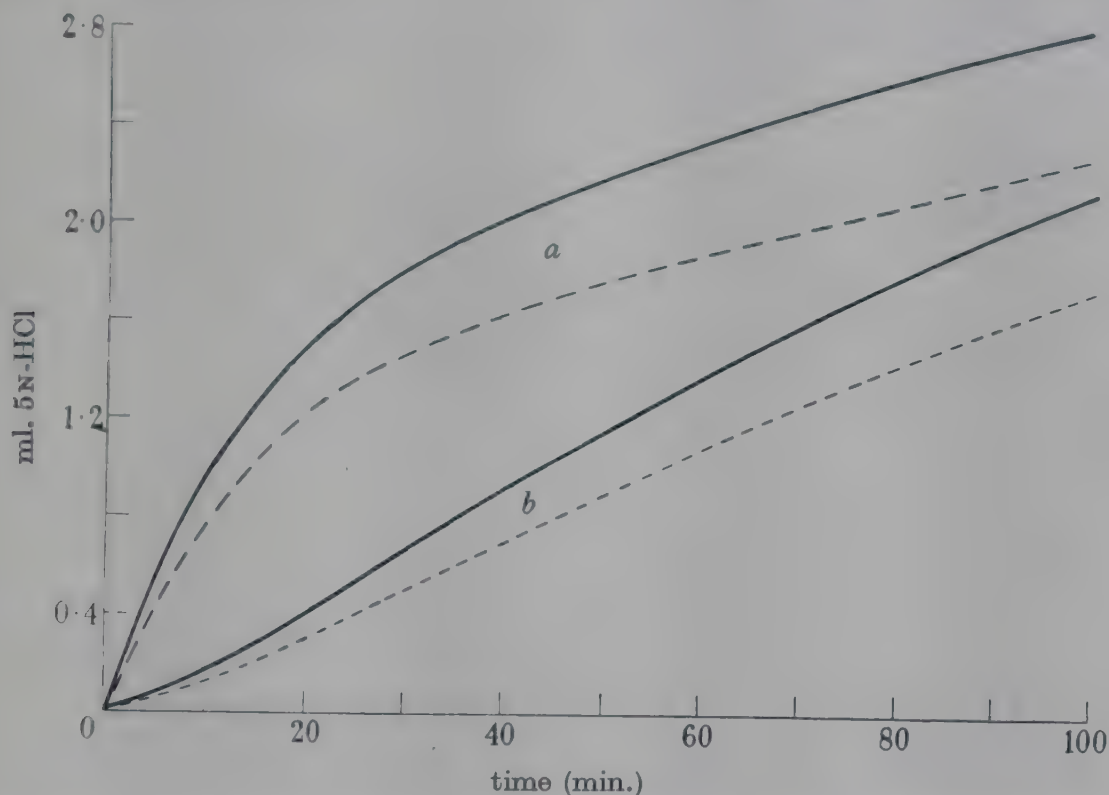


FIGURE 1. Titre against reaction time curves for 2.0 g. sugar in 2N alkali at 45° C.
a, fructose; b, glucose.

results the logarithmic plots of figure 2 are constructed. It will be noticed that in the later stages of the reaction a first-order law is followed both with glucose and fructose. Moreover, the final slopes of the curves *a* and *b* in figure 2 are identical. However, the initial first-order constants for the two sugars are different, that for fructose being the greater. Figure 2 indicates that in addition to the alkaline degradation, mutual interconversion of the sugars is occurring. Since fructose degrades more rapidly than glucose, interconversion causes the apparent first-order constant of glucose to increase steadily in the early stages of the reaction, while the reverse applies to fructose. A mixture of sugars is eventually obtained with a stationary value of the ratio glucose: fructose. From this point onwards the curves *a* and *b* of figure 2 are rectilinear. Essentially similar results were obtained with the other alkali concentrations.

The initial ratio saccharinic acid:lactic acid is greater for glucose than for fructose, the values being 0.50 and 0.22 respectively, and these ratios are independent of hydroxyl concentration and temperature in the range studied. As interconversion

proceeds during a run, the ratio falls with glucose and rises with fructose, as would be expected. These points are illustrated in figure 1.

The results, based on some 300 observations, are summarized in table 1. From the initial slopes of the logarithmic plots the first-order constants (c and d respectively in table 1) are obtained for glucose and fructose, the final slope giving the first-order constant β_2 . The intercepts of the final straight lines on the $t = 0$ axis are denoted by I_G and I_F .

The logarithmic plots, which are all similar to figure 2, suggest that all the reactions occurring in alkaline solution are first-order with respect to the sugar concentration. This was confirmed in separate experiments in which it was shown that the initial rates of degradation are proportional to the sugar concentration.

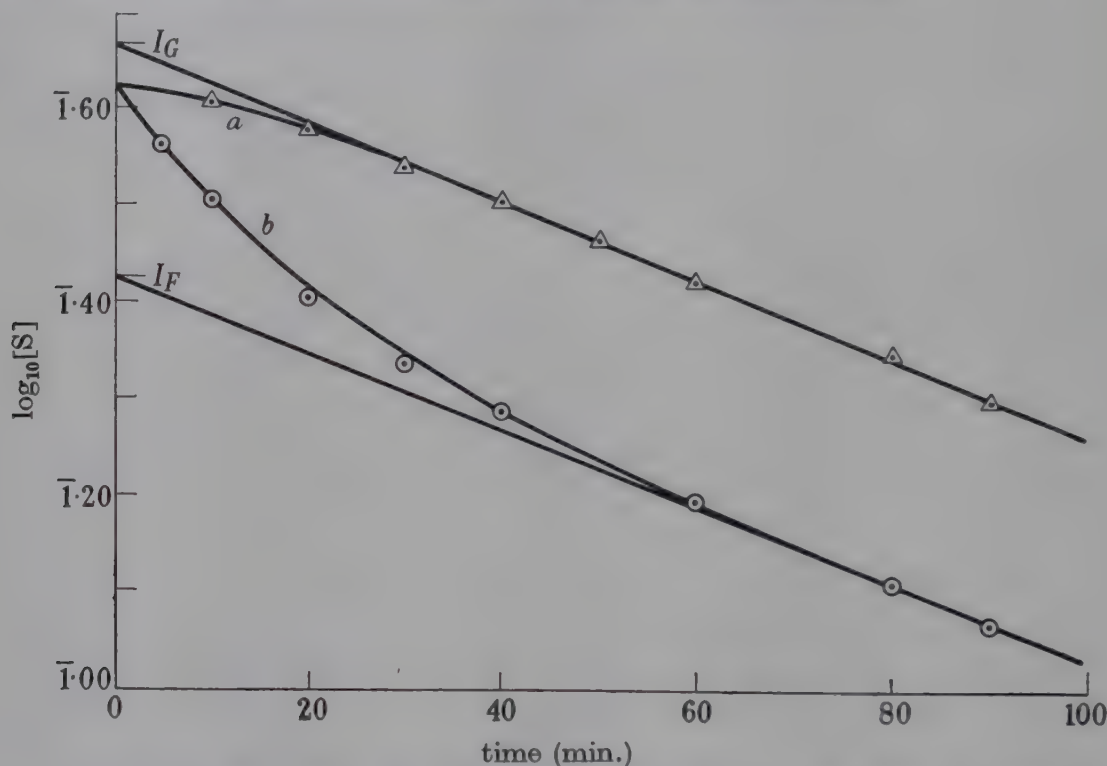


FIGURE 2. Plot of $\log_{10}[S]$ against t for 2.0 g. sugar in 2N alkali at 45° C. $[S]$ = concentration of sugar. a , glucose, $\triangle$; b , fructose, $\odot$. — calculated curves.

KINETICS

In this section the kinetics will be worked out on the assumption that all the reactions are first-order in the sugar concentration; the application to the experimental results will be subsequently investigated.

The simplest formal mechanism is



where a, b are the first-order constants for the interconversion reactions and c, d those for the degradations. If $[S_G]$, $[S_F]$ are the concentrations of glucose and fructose respectively, then the two equations below follow immediately:

$$-d[S_G]/dt = (a + c)[S_G] - b[S_F], \quad (2)$$

$$-d[S_F]/dt = -a[S_G] + (b + d)[S_F]. \quad (3)$$

The use of these equations involves the assumption that the concentrations of any intermediates which may be present in the above reactions change at a rate which is small compared to $d[S_G]/dt$ and $d[S_F]/dt$. Thus equations (2) and (3) will not apply to the very early stages of the reaction; this is not likely to be of any practical importance however, since the times concerned will only be of the order of a few seconds (see part I, section B).

From equation (2)

$$[S_F] = \{(a+c)[S_G] + d[S_G]/dt\}/b, \quad (4)$$

and elimination of $[S_F]$ from equations (3) and (4) gives the equation

$$d^2[S_G]/dt^2 + (a+b+c+d)d[S_G]/dt + \{ad+c(b+d)\}[S_G] = 0. \quad (5)$$

The solution of this is

$$[S_G] = A e^{\beta_1 t} + B e^{\beta_2 t}, \quad (6)$$

where β_1, β_2 are the roots of

$$\beta + \frac{ad+c(b+d)}{\beta} = -(a+b+c+d), \quad (7)$$

and A, B are two arbitrary constants whose values are determined by the initial conditions. Combination of equations (4) and (6) gives

$$[S_F] = \{A(a+c+\beta_1)e^{\beta_1 t} + B(a+c+\beta_2)e^{\beta_2 t}\}/b. \quad (8)$$

From equations (6) and (8) it follows that

$$[S_F] + [S_G] = -\{A(d+\beta_2)e^{\beta_1 t} + B(d+\beta_1)e^{\beta_2 t}\}/b, \quad (9)$$

since

$$\beta_1 + \beta_2 = -(a+b+c+d).$$

Two sets of initial conditions are important in practice.

(I) Fructose alone present initially; i.e. $[S_G] = 0$ and $[S_F] = [S_F]_0$ for $t = 0$. For these boundary conditions $A = -B = b[S_F]_0/(\beta_1 - \beta_2)$ and equation (9) becomes

$$[S_F] + [S_G] = [S_F]_0 \{(d+\beta_2)e^{\beta_1 t} - (d+\beta_1)e^{\beta_2 t}\}/(\beta_2 - \beta_1). \quad (10)$$

(II) Glucose alone present initially; i.e. $[S_F] = 0$ and $[S_G] = [S_G]_0$ for $t = 0$. For these initial conditions

$$A = [S_G]_0(a+c+\beta_2)/(\beta_2 - \beta_1),$$

$$B = -[S_G]_0(a+c+\beta_1)/(\beta_2 - \beta_1),$$

and equation (9) becomes

$$[S_F] + [S_G] = [S_G]_0 \{(c+\beta_2)e^{\beta_1 t} - (c+\beta_1)e^{\beta_2 t}\}/(\beta_2 - \beta_1). \quad (11)$$

Equations (10) and (11) may be applied directly to the experimental results described earlier, since $[S_F] + [S_G] \equiv [S]$ is the total residual sugar concentration at time t . The constants c, d and β_2 are obtained as already described. (It is assumed that β_2 is the numerically smaller root of equation (7).) The constant β_1 may then be calculated from equations (10) or (11). For this purpose it is simplest to use the intercepts I_G and I_F , which are given by

$$I_G = \log \left\{ \frac{[S_G]_0}{\beta_1 - \beta_2} (c + \beta_1) \right\}, \quad I_F = \log \left\{ \frac{[S_F]_0}{\beta_1 - \beta_2} (d + \beta_1) \right\}. \quad (12)$$

From the values of β_1, β_2, c, d , the values of a, b may be calculated from equation (7). They are given in table 1. No great accuracy is claimed for these velocity constants; they are probably reliable to 10 %.

TABLE 1

KOH, N	45° C					25° C	
	1	2	2.84	4	5	1	5
$-10^2\beta_1, \text{hr.}^{-1}$	213	408	453	600	744	11.4	43.2
$-10^2\beta_2, \text{hr.}^{-1}$	38.2	55.3	60.6	70.5	68.9	2.13	4.1
$I_a(\log_{10})$	1.670	1.663	1.660	1.653	1.648	1.632	1.67
$I_F(\log_{10})$	1.568	1.426	1.335	1.209	1.160	1.545	1.188
$10^2a, \text{hr.}^{-1}$	90	72	59	54	41	4.3	1.8
$10^2b, \text{hr.}^{-1}$	82	174	174	185	220	4.5	13.6
$10^2c, \text{hr.}^{-1}$	20	25	30	36	41	1.1	2.9
$10^2d, \text{hr.}^{-1}$	60	186	252	396	510	3.7	29

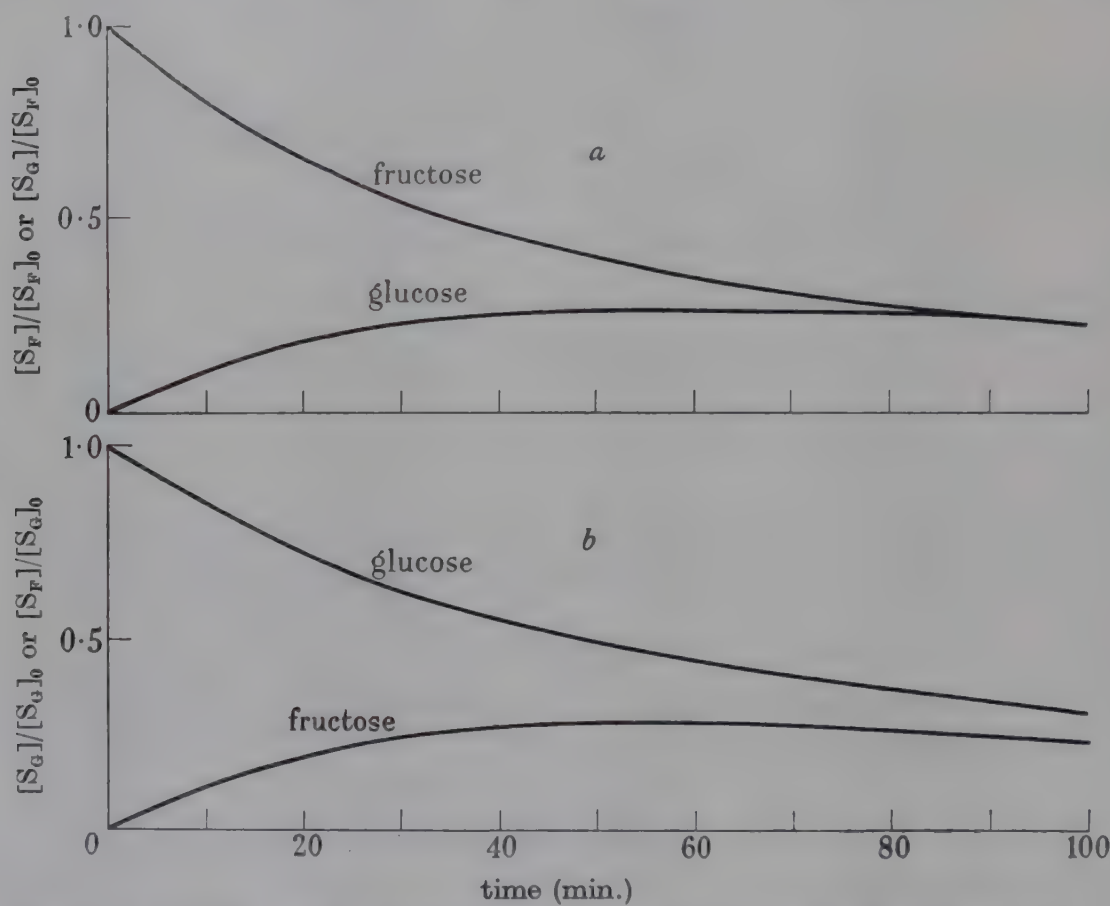


FIGURE 3. Course of the interconversion and degradation in 1N alkali at 45° C.
a, fructose initially; b, glucose initially.

In order to obtain an indication of the applicability of the kinetic scheme, it is desirable to compare the values of $\log [S]$ obtained experimentally with those calculated from equations (10), (11). This has been done in figure 2 for 2N alkali at 45° C. The curves represent the values calculated from equations (10) and (11) using the constants given in table 1. The experimental points are also shown and the agreement is very satisfactory. Similar results are obtained for the other alkali concentrations at 45° C, and also for those at 25° C. Thus the experimental results can be accounted for satisfactorily on the basis of scheme (1).

The courses of the reactions in 1 N and 5 N alkali at 45° C are illustrated in figures 3 and 4. The concentrations of the two sugars are calculated from equations (6) and (8) and are shown in these diagrams as fractions of the initial sugar concentration.

Subsidiary experiments showed that the rates of degradation and interconversion exhibited no observable change on increasing the total electrolyte concentration from 5.00 to 6.22 N, or on substituting NaOH for KOH. Thus a direct comparison may be made between the data of part I and the present paper.

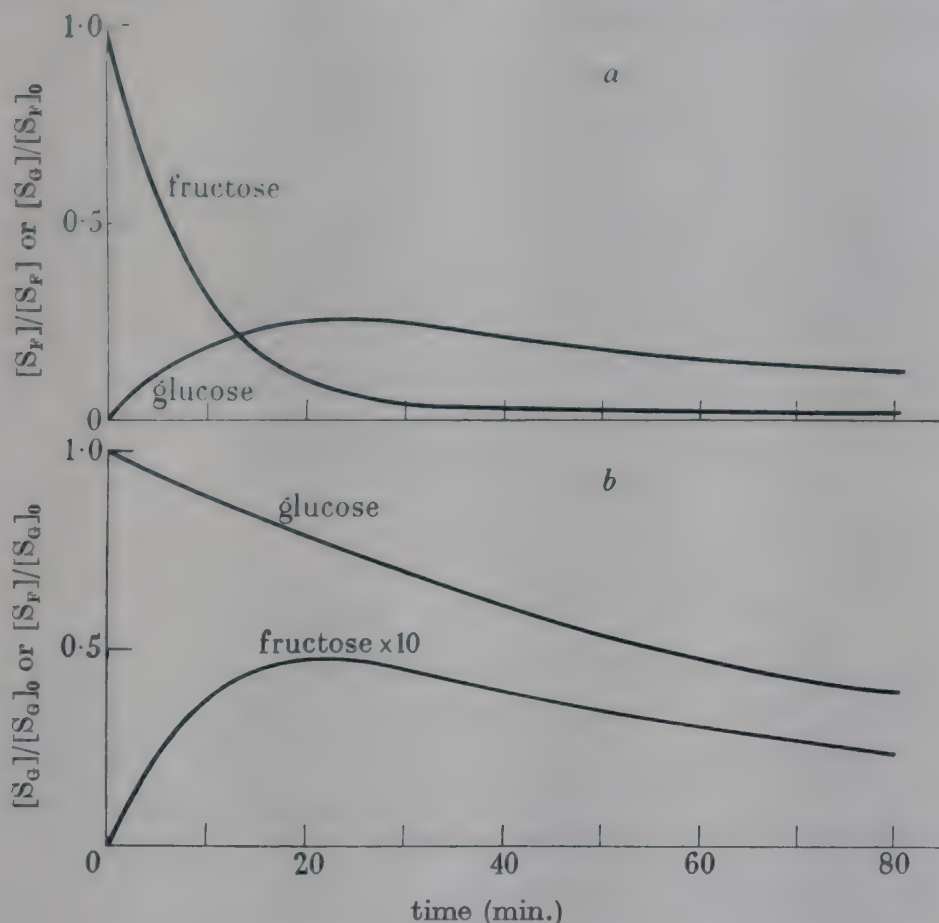


FIGURE 4. Course of the interconversion and degradation in 5 N alkali at 45° C.
a, fructose initially; b, glucose initially.

GENERAL DISCUSSION

(1) *Correlation of the oxidation, degradation and interconversion results and evaluation of the absolute velocity constants*

The degradation and interconversion of glucose and fructose have been shown to be of first-order with respect to $[S]$. This is consistent with a mechanism in which single rate-determining steps involve intermediates having concentrations proportional to $[S]$. In the case of glucose, the concentrations of the autooxidation intermediates XH_2^- and XH^- are proportional to $[S]$ (see part I); it has also been shown that degradation and interconversion are suppressed by the presence of oxygen. The assumption that the rate-determining steps in degradation and interconversion involve these intermediates (or others which are in mobile equilibrium with them) is therefore plausible. Confirmation of this mechanism is provided by the

fact that it can account satisfactorily for the dependence of the rates of degradation and interconversion on $[\text{OH}^-]$ as shown below.

Assuming this mechanism, we may write, when the fructose concentration is zero (see part I, equations (V) to (VIII))

$$-d[\text{S}_\text{G}]/dt = k_5[\text{XH}_2^-][\text{H}_2\text{O}] + k_6[\text{XH}^-][\text{H}_2\text{O}] + d[\text{S}_\text{F}]/dt, \quad (13)$$

$$d[\text{S}_\text{F}]/dt = k_7[\text{XH}_2^-][\text{H}_2\text{O}] + k_8[\text{XH}^-][\text{H}_2\text{O}]. \quad (14)$$

From equations (2), (3), (13) and (14),

$$a = [\text{H}_2\text{O}]\{k_7[\text{XH}_2^-] + k_8[\text{XH}^-]\}/[\text{S}_\text{G}], \quad (15)$$

$$c = [\text{H}_2\text{O}]\{k_5[\text{XH}_2^-] + k_6[\text{XH}^-]\}/[\text{S}_\text{G}]. \quad (16)$$

By substituting the values of $[\text{XH}_2^-]$, $[\text{XH}^-]$ given in equations (10) and (13) of part I in equations (15) and (16) above, we obtain

$$a = p \cdot [\text{GH}_2][\text{OH}^-]\{k_7 + \alpha_3 k_8[\text{OH}^-]\}/[\text{S}_\text{G}], \quad (17)$$

$$c = p \cdot [\text{GH}_2][\text{OH}^-]\{k_5 + \alpha_3 k_6[\text{OH}^-]\}/[\text{S}_\text{G}], \quad (18)$$

where

$$p = \frac{\alpha_1 k_2 [\text{H}_2\text{O}]\{[\text{OH}^-] + k_1/(\alpha_1 k_2)\}}{\alpha_3 k_4 r_2\{[\text{OH}^-] + k_3 r_1/(\alpha_3 k_4 r_2)\}}. \quad (19)$$

From the results of part I it follows that at 25°C,

$$k_1/(\alpha_1 k_2) = 0.259 \text{ mole l.}^{-1}, \quad k_3/k_4 = 0.278$$

and $r_1 \approx r_2$. Reasons have also been given for believing that $\alpha_3 \approx 1 \text{ l. mole}^{-1}$. Thus at 25°C the ratio of the terms in brackets in equation (19) is unity to within 10% for all OH^- concentrations studied and p is nearly independent of $[\text{OH}^-]$. This will also be true at 45°C, since both terms involving velocity constants have small negative temperature coefficients and are therefore always smaller than $[\text{OH}^-]$ in the range studied. Hence we may write approximately

$$p = \alpha_1 k_2 [\text{H}_2\text{O}]/(\alpha_3 k_4 r_2). \quad (20)$$

According to equations (17) and (18), $a[\text{S}_\text{G}]/([\text{GH}_2][\text{OH}^-])$ and $c[\text{S}_\text{G}]/([\text{GH}_2][\text{OH}^-])$ should therefore be linear in $[\text{OH}^-]$. Figures 5*a* and *b* shows that this is so at 45°, within experimental error.

From the slopes and intercepts of the lines, the values of $p\alpha_3 k_8$, $p\alpha_3 k_6$, pk_7 and pk_5 may be estimated. Corresponding values at 25°C may be deduced from the three experimental determinations available. Since $p\alpha_3/[\text{H}_2\text{O}]$ may be calculated directly from equation (20) (using the values given in tables 8 and 10 in part I), these results enable $k_8[\text{H}_2\text{O}]$, $k_6[\text{H}_2\text{O}]$, $k_7[\text{H}_2\text{O}]/\alpha_3$ and $k_5[\text{H}_2\text{O}]/\alpha_3$ to be estimated. Using the values $\alpha_3 = 1 \text{ l. mole}^{-1}$ and $[\text{H}_2\text{O}] = 50 \text{ mole l.}^{-1}$ (see part I), we may evaluate k_5 and k_7 ; k'_1 and k'_2 may then be calculated from equation (14) of part I.

The values obtained for these velocity constants, together with those for k_1 , k_2 , k_3 , k_4 obtained in parts I and II, are given in table 2.

No great accuracy can be claimed for the values of the velocity constants given in table 2, and consequently the frequency factors can only be considered to indicate orders of magnitude. It would appear, however, that A_1 and A_2 are appreciably higher

than the normal value, while most of the others are rather low. The high values of A_1 and A_2 may be due in part to a loosening of the pyranose ring in the transition state. This is consistent with the low value found for A'_2 , since this refers to a reaction

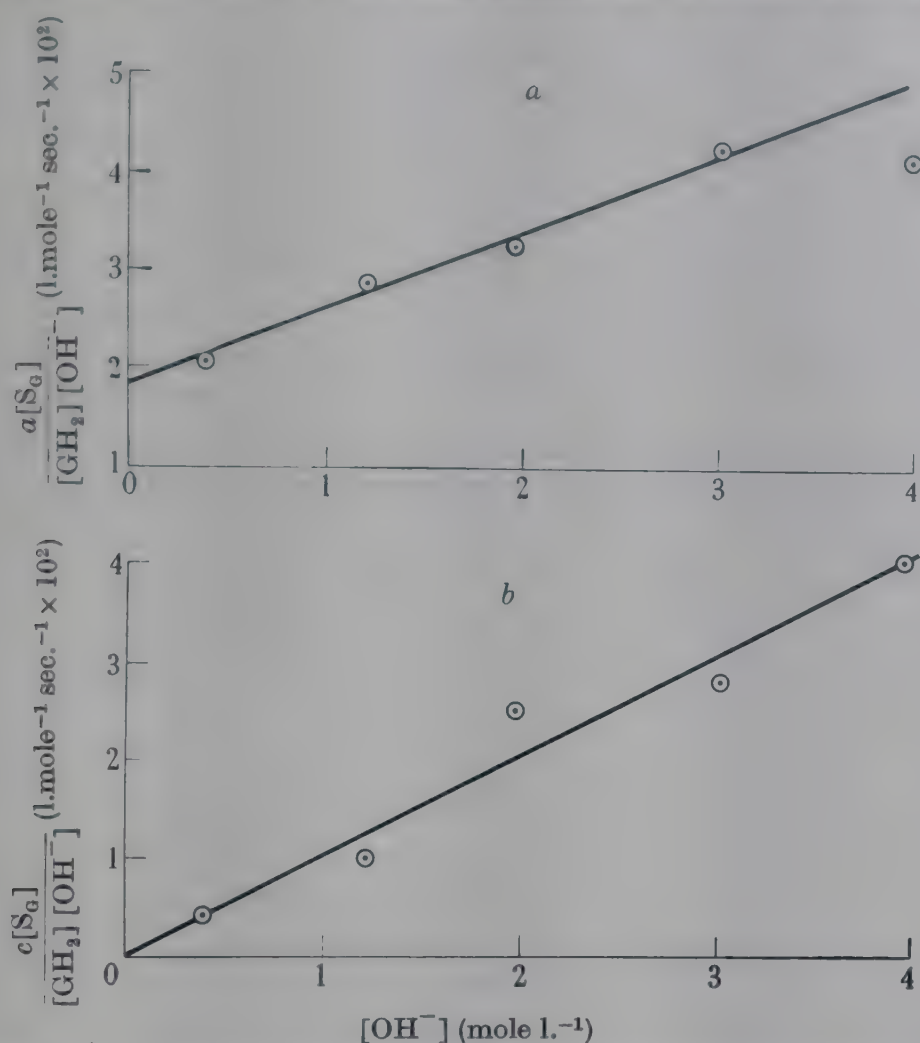


FIGURE 5. *a.* Relation of the rate of interconversion to the hydroxyl concentration. *b.* Relation of the rate of degradation to the hydroxyl concentration.

TABLE 2. VELOCITY CONSTANTS, L.MOLE⁻¹ SEC.⁻¹

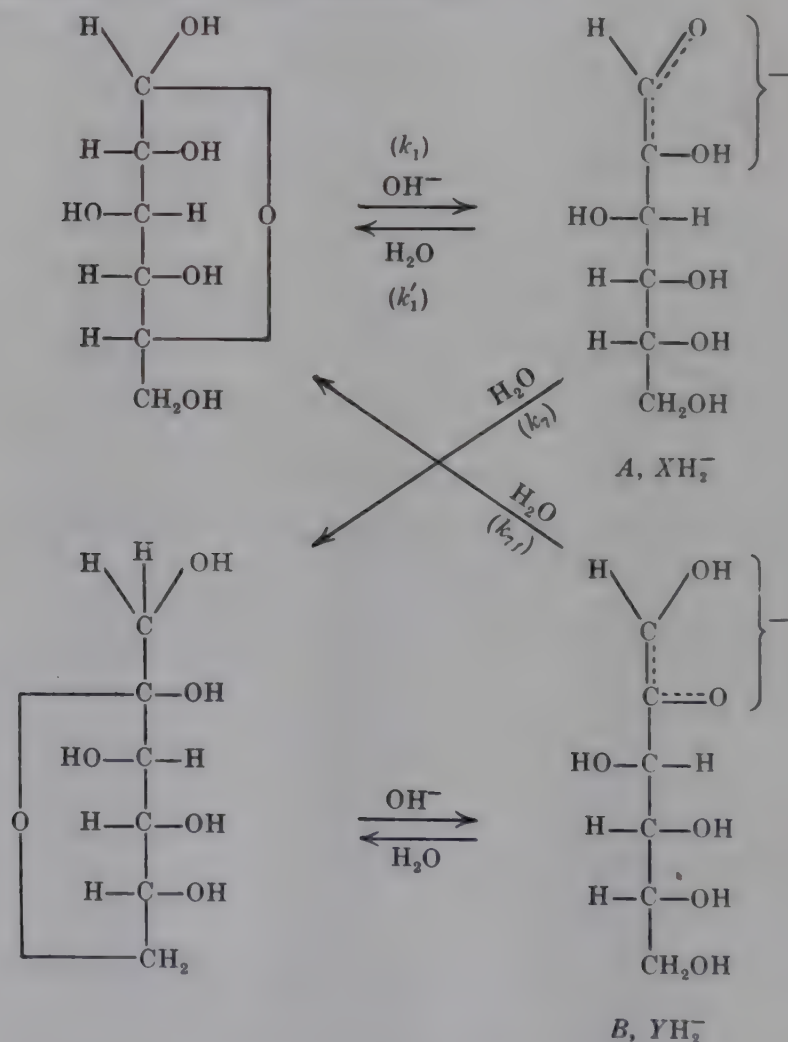
	15° C	25° C	35° C	45° C	<i>E</i> (kcal.)	<i>A</i>
k_1	—	1.79×10^{-3}	7.39×10^{-3}	—	25.9	2×10^{16}
k'_1	—	$\leq 10^{-4}$	—	$\leq 10^{-3}$	—	—
k_2	—	8.65×10^{-5}	4.12×10^{-4}	—	28.5	6×10^{16}
k'_2	—	1.9×10^{-3}	—	6.8×10^{-3}	11.8	9×10^5
k_3	2.0×10^3	3.0×10^3	—	—	6.9	3×10^8
k_4	6.0×10^3	10.8×10^3	—	—	10.0	2×10^{11}
k_5	—	$< 2 \times 10^{-5}$	—	$< 10^{-4}$	—	—
k_6	—	2.4×10^{-4}	—	9.5×10^{-4}	12.8	6×10^5
k_7	—	6.4×10^{-4}	—	1.7×10^{-3}	9.2	4×10^3
k_8	—	2.4×10^{-4}	—	7.2×10^{-4}	10.5	1×10^4

in which a ring is closed. Reactions (VII), (VIII) also lead to ring closure, and have small frequency factors. A further factor which must be considered is the change in the distribution of charge in the transition state. For example, in the reactions of XH_2^- with H_2O it is reasonable to assume that the greater portion of the negative charge, which was initially delocalized in the XH_2^- ion, is carried by the incipient

OH^- ion in the transition state $(\text{XH}_2\cdots\text{H}\cdots\text{OH})^-$. This would result in a decrease in entropy and a low frequency factor. A similar effect may operate in the reactions involving XH_2^- or XH^- and oxygen.

(2) The nature of the intermediates

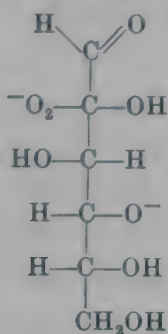
It is clear that neither of the primary rate-determining steps for the reactions discussed in this paper are the same as that in the muta-rotation reaction in the presence of alkali, since the latter process occurs very much more rapidly under comparable conditions. Thus a simple proton transfer from H_2O to the ether oxygen of GH^- is excluded as a rate-determining step, although as mentioned previously this would be satisfactory on purely kinetic grounds. The most reasonable reaction would seem to be the ionization of a $\text{C}-\text{H}$ bond; the proton which would be detached most readily is that on C_2 , since the resulting ion would be stabilized by mesomerism. Further, since methyl glucoside is unreactive in the presence of alkali, it must be assumed that formation of XH_2^- involves the opening of the pyranose ring. The intermediate XH_2^- is therefore considered to have the structure *A*.



This structure is clearly consistent with the nature of the oxidation products of glucose. The reaction between XH_2^- and oxygen presumably produces the peroxide

ion $\begin{array}{c} \text{H} \quad \text{O} \\ \diagdown \quad \diagup \\ \text{C} \\ | \\ \text{O}_2-\text{C}-\text{OH} \end{array}$, which subsequently breaks between C_1 and C_2 . The scission of an

adjacent C—C bond is a common feature in peroxide decomposition (see e.g. Leffler 1949). The corresponding ion YH^- formed from fructose probably has the structure *B* above. This would react with oxygen to give the same oxidation products as glucose. The ion XHO_2^- formed by ionization of $XH_2O_2^-$ or from XH^- and oxygen will have a structure in which one OH group is ionized, e.g.



It would be expected that ring closure between the —O on C_4 and the aldehyde group would occur readily, and this could account for the formation of the comparatively stable peroxide from this ion.

By reaction with water, which here functions as an acid, *A* and *B* can both form either glucose or fructose, depending on whether a proton adds to oxygen or carbon. Thus in the case of structure *A*, addition to oxygen will result in formation of fructose, while addition to carbon (C_2) will give glucose. As might be anticipated addition to oxygen appears to occur more readily. This is apparent from the values of k'_1 and k_7 . This mechanism for interconversion is supported by the results of table 1, which show that the rate of conversion of fructose to glucose is about equal to the rate of conversion of glucose to fructose in normal alkali, i.e. $k_{7f}[YH_2^-] \approx k_7[XH_2^-]$. Since $k_7 \gg k'_1$, $k_{7f}[YH_2^-] \gg k'_1[XH_2^-]$. It follows therefore that the conversion of fructose to glucose does not follow the course $YH_2^- \rightarrow XH_2^-$, but must proceed as shown above.

It is found experimentally that the nature of the oxidation products does not depend on the concentration of alkali present. Thus it is reasonable to assume that in XH^- , as in XH_2^- , a proton is lost from C_2 . The second proton is lost from a hydroxyl group. Since this type of ionization occurs very rapidly, it is clear that the same XH^- ion is produced by direct ionization of XH_2^- as is formed by the attack of OH^- on GH^- . From the reactions of XH^- it seems likely that the proton is lost from the hydroxyl group on C_4 (see *C*, p. 97).*

Interconversion can also occur through doubly charged ions of the type XH^{2-} , but in this case the proton addition occurs more readily to C_2 than to oxygen (cf. k'_2 and k_8). This is understandable because C_2 is now relatively more negative than in XH_2^- , being situated between two negative charges.

(3) The formation of lactic and saccharinic acids

It has been pointed out above that since the kinetic results indicate a first-order dependence on sugar concentration over the whole observable course of reaction,

* The presence of the ion *C* seems necessary to account for the formation of lactic and saccharinic acids, but the results do not exclude the existence of other doubly charged XH^{2-} ions in which hydroxyl groups other than that on C_4 are ionized. Should these ions be present, the overall kinetics will not be affected, but k'_2 , k_4 , k_8 and k_8 as defined in the text will not be absolute constants.

the formation of lactic and saccharinic acids from XH^- cannot involve the formation of any intermediate with a half-life greater than that of the reaction (VI) (approximately 1 min. at 25° C, calculated from k_6). If this were not so, the rate of formation of lactic acid would be initially zero, and would increase to a maximum as the concentration of the intermediate builds up. Experimentally it is found that lactic acid is formed at a finite rate in the earliest stages of the reaction which can be investigated. In the case of glucose there is an increase in rate, but this can be entirely accounted for by interconversion. With fructose, the rate of formation of lactic acid falls off steadily from the earliest measurements which were frequently taken after conversions as low as 5 %. Glyceraldehyde, dihydroxyacetone and pyruvic aldehyde have been suggested as intermediates in the formation of lactic acid from glucose; there appears to be general agreement that pyruvic aldehyde does participate. Unfortunately, the rates of decomposition of these compounds in alkaline solution are not known with certainty. Evans & Hass (1926) have studied the reactions of glyceraldehyde. Taken at their face value their results indicate that the substance is rather stable, having a half-life of about 10 hr. in N -KOH at room temperatures. However, it is possible that the major portion of their glyceraldehyde was present as a polymer. This also applies to the results of Bernhauer & Gorlich (1929) on the formation of lactic acid from pyruvic aldehyde, which is known to polymerize very rapidly in the presence of water. However, the monomeric aldehydes may react much more rapidly, in which case the kinetic results would not be inconsistent with their participation in the reaction. Pyruvic aldehyde osazone (see e.g. Evans, Edgar & Hoff 1926) and small amounts of the free aldehyde (see e.g. Bernhauer & Wolf 1929) have been obtained from sugar in alkaline solution, but this does not necessarily prove that the aldehyde as such is an intermediate. These results would be equally compatible with formation of an intermediate which normally reacts to give mainly lactic acid, with a little pyruvic aldehyde, but which in the presence of phenyl hydrazine gives mainly the osazone.

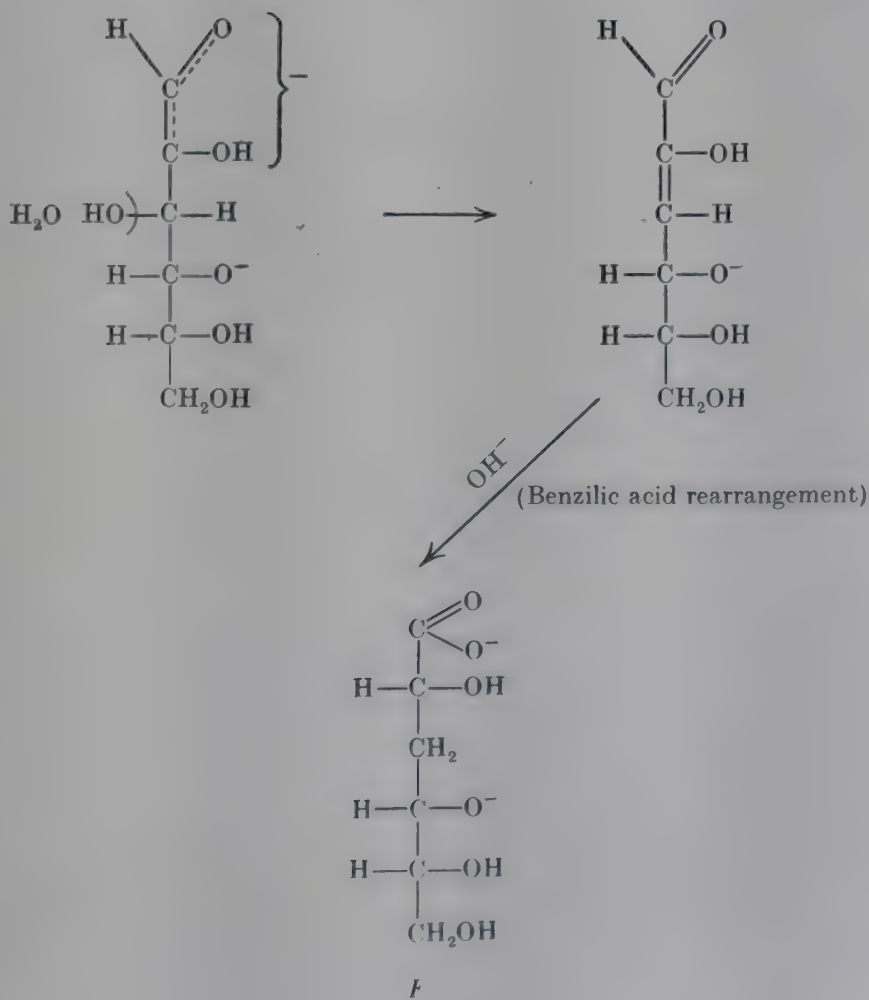
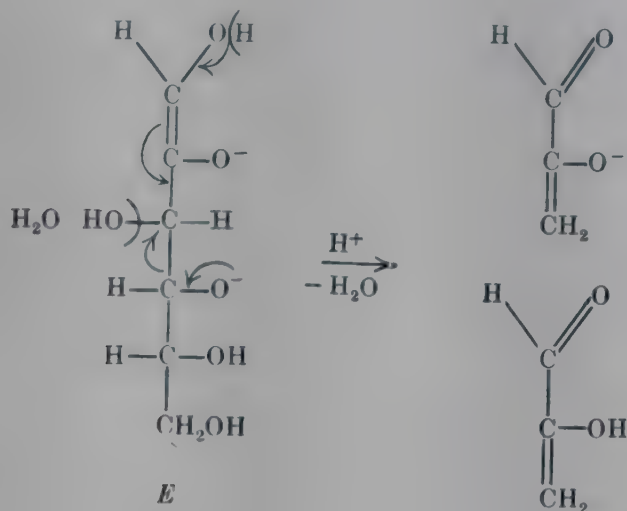
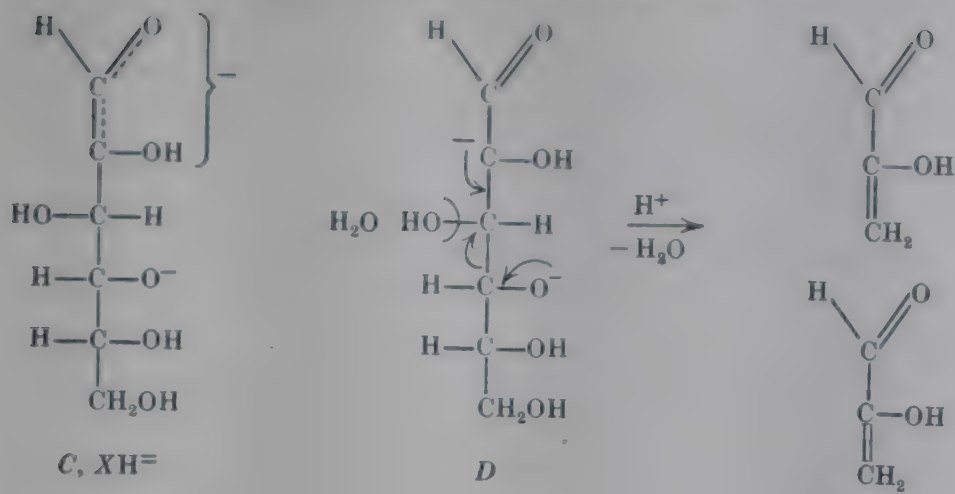
A possible mechanism is shown below in *D*. The hydroxyl group on C_3 is removed by a water molecule, and the positive charge remaining on C_3 initiates the electronic displacements shown. The driving force of the reaction may be regarded as the formation of the double bond between C_2 and C_3 , and the carbonyl group at C_4 . Similar processes for YH^- are shown in *E*.

The mechanism of the formation of saccharinic acids from glucose has been discussed by Isbell (1944). In concentrated alkali solutions (normal and higher) metasaccharinic acid is mainly formed. The production of this acid (*F*) from XH^- is shown below, the mechanism being similar to Isbell's.

A similar mechanism applies in the corresponding reaction of fructose.

(4) *The reaction in weakly alkaline solutions*

It is well known that in weakly alkaline solutions interconversion is much more rapid than degradation (see e.g. Wolfrom & Lewis 1928). Our results are in agreement with this. In dilute alkaline solutions where $[XH_2^-] \gg [XH^-]$, the relative rates of formation of fructose and lactic acid from glucose are proportional to k_7 and k_5



respectively (equations (15), (16)). The former rate should therefore be greater by at least a factor of 30 at 25° C according to the results given in table 2.

The rate of conversion of glucose to fructose has been studied polarimetrically by Wolfrom & Lewis (1928) at 35° C in lime-water solution (initial pH 10.6). By drastic extrapolation of the results of figure 5*a* we may estimate that under these conditions $a = 1.8 \times 10^{-2} \text{ hr.}^{-1}$, i.e. the initial rate of conversion of glucose is 1.8 % per hr. Wolfrom & Lewis's figures lead to a value of 0.5 % per hr. The agreement is probably as good as could be expected. Our results refer to solutions in which the total electrolyte concentration is 5*N*; it has been observed that the addition of sodium bromide increases the rate. Further, it is not clear under what conditions the pH recorded by Wolfrom & Lewis was measured.

The authors wish to express their thanks to Mr B. C. Miller and Miss J. R. Pearl for experimental assistance.

REFERENCES

- Bernhauer, K. & Görlich, B. 1929 *Biochem. Z.* **212**, 452.
 Bernhauer, K. & Wolf, H. 1929 *Biochem. Z.* **212**, 443.
 Evans, W. L., Edgar, R. H. & Hoff, G. P. 1926 *J. Amer. Chem. Soc.* **48**, 2665.
 Evans, W. L. & Hass, H. B. 1926 *J. Amer. Chem. Soc.* **48**, 2703.
 Isbell, H. S. 1944 *Bur. Stand. J. Res., Wash.*, **32**, 45.
 Leffler, J. E. 1949 *Chem. Rev.* **45**, 385.
 Nef, J. U. 1910 *Liebigs Ann.* **376**, 1.
 Shaffer, P. A. & Friedemann, T. E. 1930 *J. Biol. Chem.* **86**, 345.
 Upson, F. W. 1911 *Amer. Chem. J.* **45**, 458.
 Wolfrom, M. L. & Lewis, W. L. 1928 *J. Amer. Chem. Soc.* **50**, 837.

The thermal conductivity of tin, mercury, indium and tantalum at liquid helium temperatures

BY J. K. HULM

The Royal Society Mond Laboratory, University of Cambridge

(Communicated by Sir Lawrence Bragg, F.R.S.—Received 20 May 1950—
 Revised 28 July 1950)

An account is given of a method of employing differential gas thermometers to measure the thermal conductivity of metals at low temperatures, with temperature differences of only 0.02° K in the specimen. Experimental curves are presented showing the variation of thermal conductivity with temperature between 1.7 and 4.3° K for pure tin, alloys of tin with mercury, pure mercury, alloys of mercury with cadmium and indium, pure indium and pure tantalum, in both superconducting and normal states.

The normal conductivity of strain-free specimens containing less than about 0.1 % of impurity appears to be mainly electronic, and to behave roughly in accordance with the theory of Wilson, although notable discrepancies arise in the detailed application of this theory. The ratio of superconducting to normal thermal conductivity varies with temperature roughly in the manner suggested by Heisenberg when the electrons are mainly scattered by

impurities, but follows a radically different curve, for which no theoretical explanation is yet available, when lattice vibrations are the dominant scattering mechanism.

For impurity contents greater than 0.1 %, or severe internal strains, the normal thermal conductivity contains an appreciable lattice component. The behaviour of the superconducting curve suggests that where crystal boundaries scatter the lattice waves, the lattice conductivity is unaltered as the metal passes from the superconducting to the normal state, but where electrons scatter the lattice waves, the lattice conductivity is reduced in this transition, possibly because of the increase in the number of scattering centres.

1. INTRODUCTION

Comparatively little quantitative data have hitherto been available on the thermal conductivity of metals at liquid helium temperatures. Previous investigations in the field, of which by far the most extensive were those of de Haas and his co-workers (1931, 1936, 1940) at Leiden, were chiefly concerned with the changes in thermal conductivity that occur during the creation or destruction of the electrically superconducting state. These experiments showed that if a pure metal cooled well below its superconducting transition temperature is subjected to an increasing magnetic field, the transition from superconducting to normal state, which occurs at the critical field strength, is accompanied by an increase in thermal conductivity, or in other words the thermal conductivity is smaller in the superconducting state than in the normal state. The measurements were not, however, sufficiently accurate to give a clear picture of the variation of thermal conductivity in the immediate neighbourhood of the superconducting transition temperature, and, moreover, did not allow a rigorous test of the theoretical expression for the thermal conductivity of a normal metal derived by Wilson (1937) from the electron theory of metals. The importance of these problems made it recently seem worth while to try to improve the technique of low-temperature thermal conductivity measurements so that a new and more accurate survey could be carried out for both superconductors and normal metals. A brief account has already been given (Hulm 1949) of the successful development of an improved apparatus, and of the preliminary results obtained for specimens of pure tin, alloys of tin with mercury and pure tantalum. These experiments are described in detail in the present paper, and additional thermal conductivity data are presented for pure mercury, alloys of mercury with cadmium and indium, and pure indium.

2. EXPERIMENTAL DETAILS

2.1. *Method*

The measurement of thermal conductivity was carried out with an apparatus similar to that used by de Haas & Rademakers (1940), in which a known heat current, produced electrically, was passed along a uniform rod of the material under investigation, and the temperature difference between two separate points on the rod determined with a pair of helium gas thermometers. The novel feature of the present experiments was the use of a miniature butyl phthalate oil manometer, read with a travelling microscope, to measure the difference in gas pressure δp existing between the two thermometer bulbs due to their temperature difference δT . Provided that

the bulbs were at the same temperature T and the same gas pressure p before a heat current was set up in the specimen rod, δT could be determined from the expression

$$\delta T = T \delta p / p. \quad (1)$$

Corrections for departure from ideal gas behaviour and for the dead space of the manometer system will be discussed in §2.4. By working with a pressure difference in the neighbourhood of 1 mm. of oil, a temperature difference as small as 0.02°K could be measured with a probable error of less than 1 % over most of the liquid helium range of temperatures, and thermal conductivity could then be determined with similar precision from δT , the power input W and the dimensions of the specimen. Since the values of δT employed were much smaller than in previous investigations, the drawback of having thermal conductivity data averaged over a range of temperature comparable with the absolute temperature of the specimen was entirely removed, and regions of rapid variation of conductivity with temperature could be surveyed with greater reliability and accuracy than before.

2.2. *Design of cryostat*

For compact arrangement of the low-temperature portion of the apparatus, the gas thermometers were made in the form of cylindrical annuli mounted coaxially with the specimen, as shown in cross-section in figure 1*a*. Both specimen and thermometers were enclosed in a highly evacuated cylindrical container, and this was in turn situated in a Dewar vessel which could be filled with liquid helium. The lower end of the specimen rod carried the electrical heater coil, while the upper end of the rod was soldered through a collar in the top of the container, thus making direct contact with the liquid helium.

The experimental conditions ensured that a minimum fraction of the total heat power generated by the heater coil was transported to the liquid helium by paths other than that provided by the specimen itself. Thus the gas pressure in the container was sufficiently low to make gas conduction unimportant, while radiation losses could be neglected at the temperatures involved. Heat losses along the heater coil current and potential leads, which were respectively 30 and 44 s.w.g. manganin wires of about the same length as the specimen, led out of the container through kovar metal-to-glass seals at the top, were completely negligible. Heat losses along the 1 mm. internal diameter copper-nickel alloy capillary tubes which connected the thermometer bulbs to manometers outside the cryostat were reduced somewhat by spiralling these capillaries to increase their length, but for specimens of fairly low thermal conductivity the capillary losses were still large enough to comprise several per cent of the total power generated by the heater. In these cases, therefore, explicit allowance was made for the capillary losses by calculating them from the dimensions of the system and the thermal conductivity of the copper-nickel alloy, the latter being determined in a separate experiment with a copper-nickel alloy specimen.

Three other measures were taken to ensure convenient and accurate working. First, the end of the german-silver tube used to evacuate the specimen container was partially closed off so as to reduce the possibility of gas molecules or radiation

from the room-temperature portion of the tube impinging on the thermometer bulbs. Secondly, the manganin wire heater coil was wound non-inductively on a radially slotted copper drum, thus ensuring that heat impulses due to sudden changes in the longitudinal magnetic field applied to the specimen were not large enough to cause

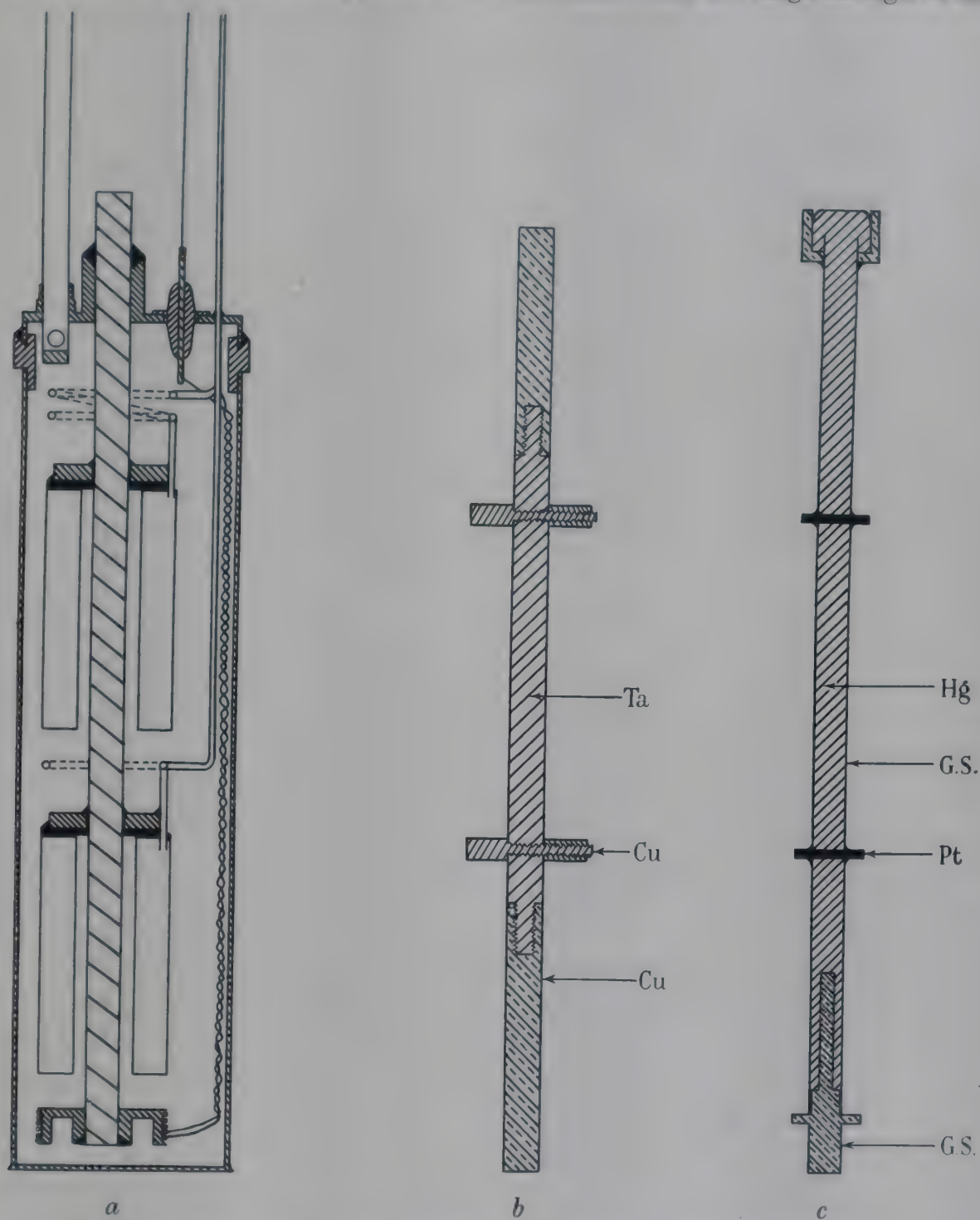


FIGURE 1. *a*, specimen holder; *b*, tantalum specimen; *c* mercury specimen.

serious deflexions of the differential oil manometer. Finally, in order to avoid the danger that the heat current flowing in the copper-nickel capillary tubes, although small compared with that flowing in the specimen, would nevertheless set up a temperature gradient in each gas thermometer bulb, a thick copper half-ring was placed in position at the top of each bulb, soldered to the bulb itself, to the capillary tube at its point of departure from the bulb and to the specimen rod.

2.3. *Manometers*

Since the temperature difference between the thermometer bulbs was determined by measuring the three quantities on the right-hand side of equation (1), and T was obtained from the vapour pressure of the liquid helium bath, a single determination of δT involved, in all, three independent pressure measurements. The liquid helium vapour pressure was measured on a conventional mercury manometer at high pressures and on a continuously evacuated oil manometer at low pressures, temperatures being calculated on the Leiden 1932 scale (Keesom 1932). The pressure difference δp and absolute pressure p for the thermometer bulbs were both measured on oil manometers, so that as far as these pressures were concerned the density of the oil did not appear explicitly in the expression for δT .

In order to minimize the dead space a miniature differential manometer was made from 3 mm. diameter uniform glass tubing waxed directly on to the ends of the thermometer capillary tubes where these emerged from the top of the Dewar vessel. By illuminating the oil menisci from above and behind, the cross-wire of a travelling microscope could be set opposite the bottom of either meniscus with a reproducibility of better than 0.01 mm., so that a pressure difference of about 1 mm. of oil could be measured with an accuracy of 1 %. Only one value of δp was employed in the determination of thermal conductivity at a given temperature, since a check experiment showed that δp was strictly proportional to the heater power W for values of the former ranging from 0.05 to 10 mm. of oil. The time taken by the oil menisci to reach a steady position after setting up the heater current was about 30 sec. and independent of the thermal conductivity of the specimen for good conductors, but increased considerably for specimen conductivities less than about 0.1 W/cm./degree, reaching a value of 10 min. for the worst conductor examined.

The absolute pressure of helium in the thermometer bulbs was indicated by a continuously evacuated oil manometer, but measurements of p were necessary at only a few temperatures in the liquid helium range, since the condition of constant volume working enabled a constant value of p/T to be substituted in equation (1) at intermediate temperatures. With an absolute pressure of between 10 and 20 cm. of oil, measurements of thermal conductivity could be carried out at temperatures down to 1.7° K before condensation in the thermometer bulbs prevented further work.

2.4. *Corrections*

In deriving δT from the measured quantities δp , p and T , account had to be taken of departures of the thermometer system from ideal behaviour, which modified the right-hand side of equation (1) by a few per cent. Explicit allowance was made for

(1) The dead space of the differential manometer and capillary tubes, using the dimensions and temperatures of the component parts. The main contribution came from the region of capillary tube immersed in liquid helium, which comprised a volume of about 0.28 ml. for each thermometer when the Dewar flask was completely filled; with a thermometer bulb volume of 8.52 ml., this resulted in a 3.3 % correction to the right-hand side of equation (1). The room-temperature region on

either side of the differential manometer had a volume of about 2.3 ml., giving only 0.37 % correction with the thermometer bulbs at 4° K.

(2) The thermolecular pressure difference between the thermometer bulbs and the manometers, using the formula of Weber, Keesom & Schmidt (1936). This correction never exceeded 2 %.

(3) The deviation of helium from perfect gas behaviour, using the second and third virial coefficients measured by Keesom & Walstra (1940, 1947) and Kistemaker & Keesom (1946). Experimental conditions ensured that this correction was also less than 2 %.

In deriving the heat current flowing in the specimen from the heat generated by the heater coil, allowance was made for

(1) the heat losses in the thermometer capillary tubes, already mentioned (2.2),
 (2) the heat generated in the leads of the heater coil, of which half was assumed to reach the specimen. This was about 1 % of the total heat input to the specimen.

Finally, accurate determination of the specimen temperature depended on a knowledge of the small temperature difference between the centre of the specimen and the surface of the liquid helium bath due to the heat current flowing between these two points. This temperature difference, about 0.1° K, was calculated from observations of the change in absolute pressure in both thermometer bulbs when the heat current was set up in the specimen.

2.5. Specimens

The characteristics of all the specimens are set out in table 1. The first member of both tin and mercury series was prepared from a sample of spectroscopically pure metal, while the succeeding alloy members were obtained by adding to this metal the various amounts of impurity listed. Tin and indium specimens were prepared by casting the metals in thin-walled glass tubes which were afterwards cracked and

TABLE 1. CHARACTERISTICS OF SPECIMENS

composition			electrical data		thermal data		important terms in K			
origin	% wt. imp.	cryst. state	T_c (° K)	ρ_0/L_0	β	$\alpha \times 10^4$	K_e latt. scat.	K_e imp. scat.	K_g elec. scat.	K_g bound scat.
Sn 2 J. 2356	0.004*	(a)	3.71	0.045	0.048	3.92	×	×	—	—
Sn 3 J. 2356 + Hg	0.033	(a)	3.68	0.577	0.656	5.70	×	×	—	—
Sn 6 J. 2356 + Hg	0.33	(a)	3.72	6.72	—	—	—	×	×	—
Sn 9 J. 2356 + Hg	4.1	(c)	—	106	—	—	—	×	×	—
In 1 J. Chem. Pure	0.1*	(a)	3.40	1.47	1.38	18.9	×	×	—	—
Ta 1 H. 8017	0.1*	(b)	4.38	157	—	—	—	×	—	×
Hg 1 J. 3100	—	(a)	—	—	—	—	×	×	—	—
Hg 2 J. 3100 + Cd	0.002	(a)	—	—	0.05	240	×	×	—	—
Hg 3 J. 3100 + Cd	0.007	(a)	—	—	0.32	255	×	×	—	—
Hg 6 J. 3100 + In	0.10	(a)	—	—	1.1	—	×	×	×	—
Hg 8 J. 3100 + In	0.39	(a)	—	—	3.4	—	×	×	×	—
Hg 9 J. 3100 + In	1.19	(c)	—	—	—	—	×	×	×	—

J. = Johnson-Matthey. H. = Hilger.

* Maker's analysis. (a) Homogeneous solid solution, few large crystals.

(b) Homogeneous solid solution, many small crystals.

(c) 2-phase, few large crystals.

removed piecemeal. These specimens were held in position in the apparatus and joined to the thermometer bulbs by Wood's alloy solder. Tantalum was made into a composite specimen (figure 1*b*) with pure copper ends and pure copper screw contacts for the thermometer bulbs, the two metals having good thermal contact on cooling due to the difference in their expansion coefficients. Mercury was cast in a thin-walled german-silver tube (figure 1*c*) already soldered in position in the apparatus; contact with the thermometers was made by contraction of the mercury on to platinum rods silver-soldered across the axis of the tube and soldered to the thermometer bulbs. This technique failed for concentrations of cadmium greater than about 0.01 %, but was successful for much higher concentrations of indium.

The crystalline state of the tin and indium specimens was judged from equilibrium diagrams and etching data. For mercury no etching data were available, but the very slow cooling technique used in casting the specimens suggests that fairly large crystal grains would have been formed. The assumption of small crystal grains in the tantalum specimen is consistent with the very high residual electrical resistance and with the difficulty of making large crystals of this metal.

Since electrical measurements could not be carried out in conjunction with the thermal measurements, the residual electrical resistance and superconducting transition temperature of each specimen was determined in a separate experiment after the specimen had been removed from the thermal conductivity apparatus. For this reason no electrical data were obtained on the mercury specimens.

3. RESULTS

3.1. *Isothermal magnetic field curves*

The thermal conductivity of the specimens listed in table 1 was measured at a series of temperatures between 1.7 and 4.3° K, and for a series of longitudinal magnetic field strengths ranging from zero up to about 1500 G at each temperature. Magnetic fields were applied primarily in order to destroy superconductivity, so that the normal thermal conductivity K_n could be determined for each specimen at temperatures below its superconducting transition point, T_c . At such temperatures the variation of thermal conductivity with magnetic field for most specimens was similar to that shown for Sn 3 in figure 2. The conductivity had a value K_s independent of field for field strengths less than the critical value, changed suddenly to a value K_n as the critical value was exceeded and the specimen passed from the superconducting to the normal state, and was then practically unaltered by a large increase in the field applied to the normal metal. For tin and mercury specimens the range of field strengths over which the main part of the change in conductivity took place was found to increase with increasing amounts of impurity, varying, for example, from a few gauss in Sn 3 (0.03 % impurity) to more than 100 G in Sn 9 (4 % impurity). In spite of its apparently low chemical impurity content, the tantalum specimen had a transition range nearly equal to that for Sn 9, presumably because high internal strains in the tantalum played the same part as impurities in the tin in broadening the transition. In both Ta 1 and Sn 9 the transition curve for increasing field was separated markedly from that for decreasing field at the same temperature, but in

neither specimen could any difference be detected between the initial and final values of thermal conductivity in zero field, such as would perhaps result if appreciable amounts of magnetic flux were locked in the metal at the end of a measuring cycle.

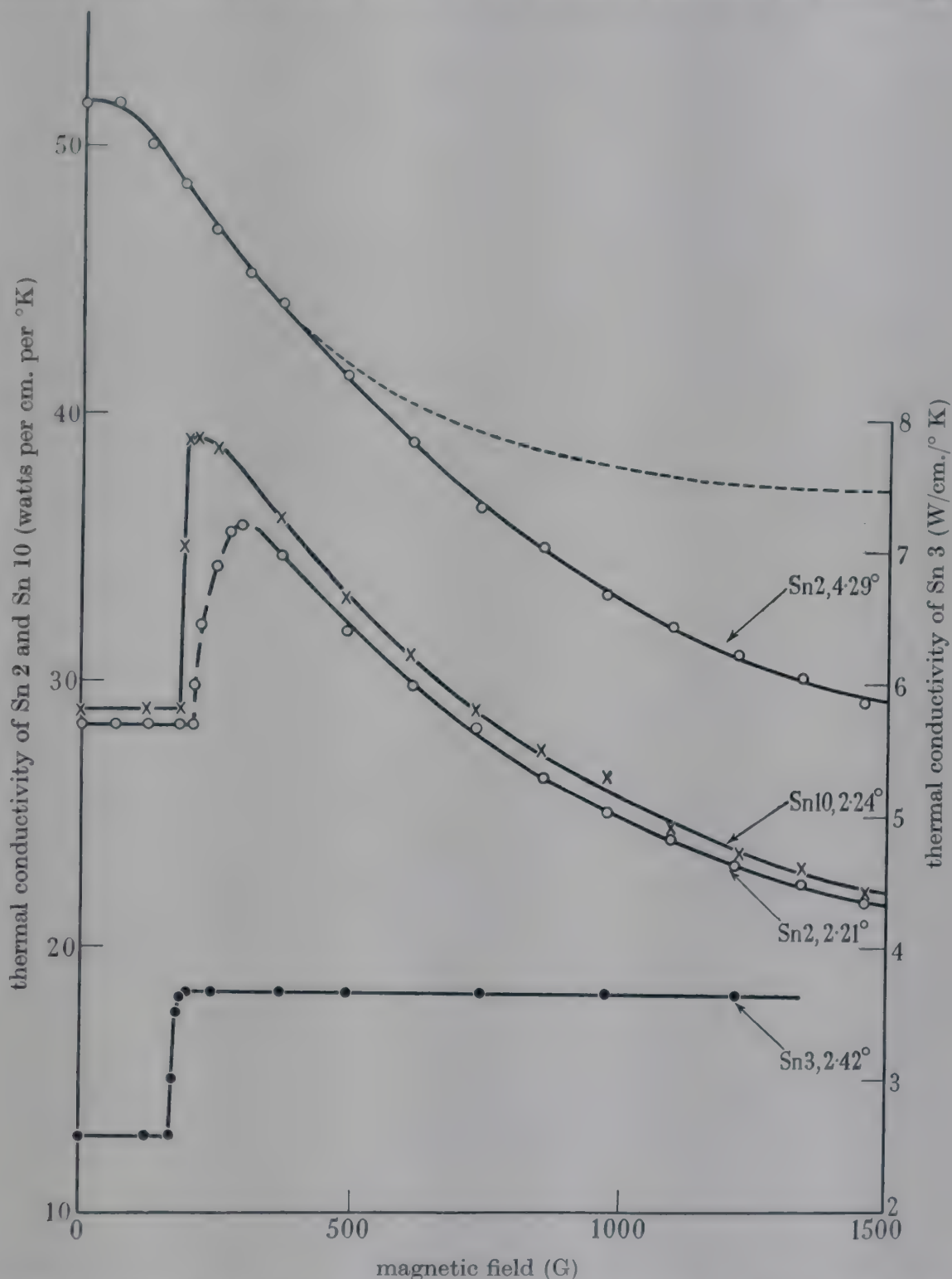


FIGURE 2. Variation of thermal conductivity of tin with magnetic field. ○ Sn 2, ● Sn 3, × Sn 10, ---- comparison of equation (2) with Sn 2 curve at 4.29° K.

As a further check K_s for each specimen in table 1 was measured at the lowest possible temperature following a run down from T_c in which magnetic field transition curves were taken at a series of different temperatures, and then remeasured at the same temperature after the specimen had been warmed up to T_c and cooled down again

in zero field. The two values of conductivity were in excellent agreement, thus indicating that if in the first cooling process there was any accumulation of locked-in flux due to a small contribution from each magnetic field cycle, this had no influence on the superconducting thermal conductivity.

The most impure mercury specimen, Hg 9, departed somewhat from the usual behaviour in that the total change of thermal conductivity with increasing magnetic field could be recognized as having two components, a fairly rapid increase of conductivity in the vicinity of the critical field value for pure mercury, and a gradual decrease of conductivity spread over the range from 0 to about 800 G. These two components could be identified with two phases which were probably mixed together in the specimen, a homogeneous solid solution of indium in mercury, showing mercury-like behaviour, and an indium-mercury intermediate compound with a very long-drawn-out transition from the superconducting to the normal state accompanied by a decrease in conductivity, such as observed by Bremmer & de Haas (1936) in lead-thallium and lead-indium alloys.

The spectroscopically pure specimen Sn 2 differed from all other specimens in table 1 in exhibiting a marked dependence of thermal conductivity on magnetic field in the normal state. Curves for this specimen in figure 2 show the steady decrease of conductivity with increasing field at 4.29° K, well above T_c , and below T_c (as 2.21° K) a similar decrease with fields greater than the critical value. In view of the high purity of Sn 2 it was surprising to find that the change in thermal conductivity associated with the transition at 2.21° K took place over a range of as much as 100 G, and, moreover, that the critical field strength for the transition was about 30 G higher than found by de Haas & Engelkes (1937) in electrical measurements on pure tin. This proved to be a spurious effect due to distortion of the magnetic field in the neighbourhood of the specimen by a superconducting ring of mercury solder which held together the two halves of the vacuum jacket. In the measurements on all specimens other than Sn 2, the mercury ring was replaced by a much thinner ring of Wood's alloy, and the distortion effects were then greatly reduced. A check experiment on a specimen Sn 10 of the same composition as Sn 2 (not listed in table 1) gave the transition curve shown at 2.24° K in figure 2, with a small transition range and a normal value of critical field.

The normal thermal conductivity of Sn 2 at temperatures below T_c was obtained by extrapolating the normal curve to zero field, using an empirical formula which represented the 4.29° K conductivity field curve quite well from 0 to about 400 G, viz.

$$\frac{K_n(0) - K_n(H)}{K_n(H)} = \frac{\lambda H^2}{1 + \mu H^2}, \quad (2)$$

where $K_n(0)$ and $K_n(H)$ are the normal conductivities in zero field and field H respectively, and λ and μ are adjustable parameters. The broken curve in figure 2 shows that equation (2) fails at high field strengths, but this did not affect the extrapolation to zero field at lower temperatures since the destruction of superconductivity was always sensibly complete at 400 G. The dependence of λ and μ on temperature could not be determined accurately in the limited range above T_c , but rough measurements indicated that both quantities were proportional to K_0^2/T^2 , and

this was assumed in the extrapolation. This is reasonable from a theoretical point of view, since Sondheimer & Wilson (1947) have derived an equation similar to (2), with λ and μ proportional to K_0^2/T^2 , for the thermal conductivity of the transition metals in transverse magnetic fields, and have pointed out that most metals may be expected on fairly general theoretical grounds to give magneto-conductivity curves of the above type.

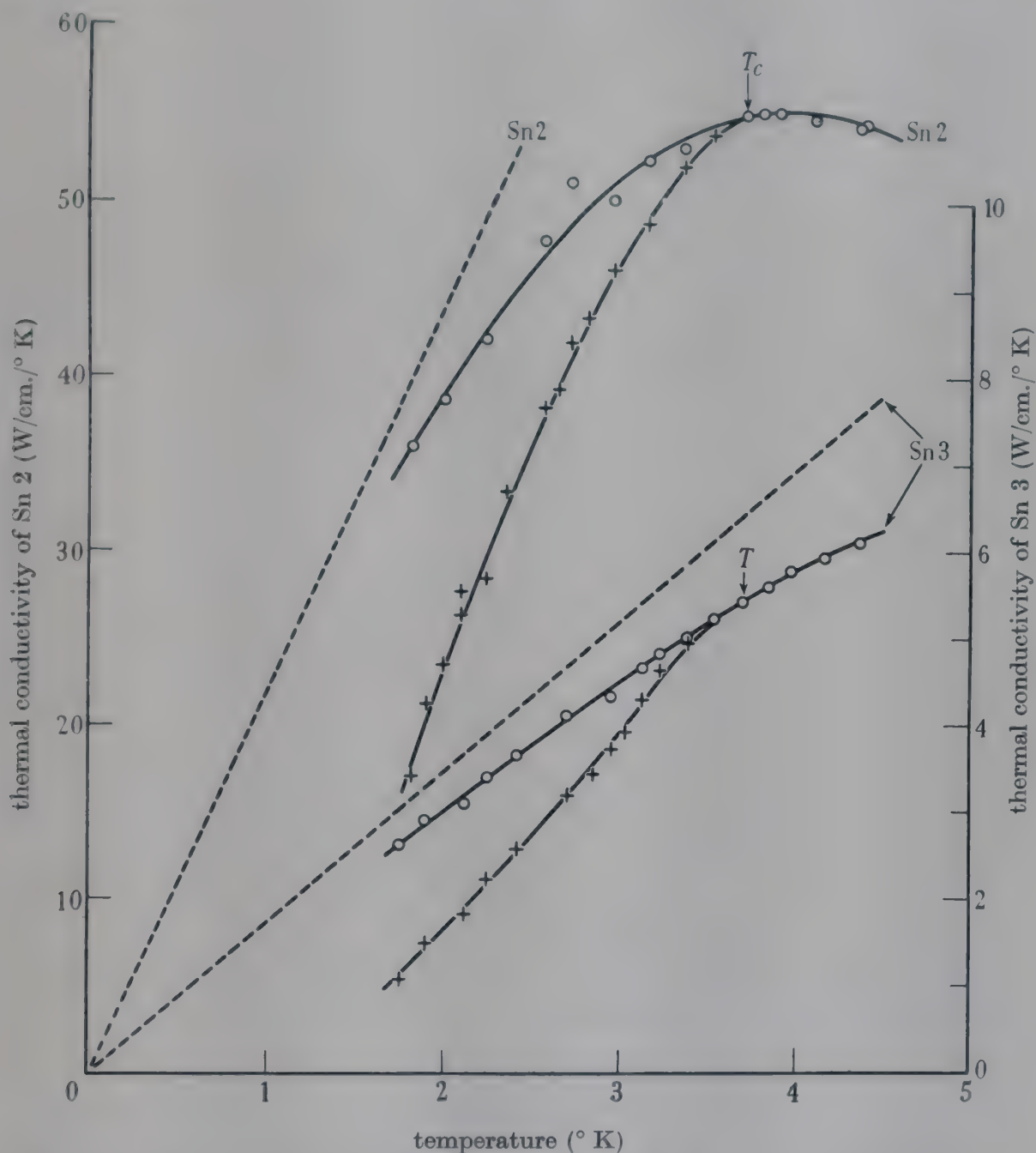


FIGURE 3. Thermal conductivity of tin specimens Sn 2 and Sn 3. $\circ$ normal state, $+$ superconducting state, ---- $L_0 T / \rho_0$.

3.2. Temperature variation of thermal conductivity

The temperature dependence of K_s and K_n for the various specimens is shown in figures 3 to 7 inclusive, the data for Hg 9 being omitted owing to the complex magnetic behaviour of this specimen. T_c marks the electrical transition temperature, where this was measured, and the dashed line represents $L_0 T / \rho_0$, where L_0 is the

Lorenz constant $\frac{1}{3}(\pi k/e)^2$ and ρ_0 is the residual electrical resistivity. Although there is a marked difference between the forms of the normal curves for tin and mercury specimens of about the same impurity content, the general behaviour is not inconsistent with that found by Gruneisen & Goens (1927) for other metals at temperatures

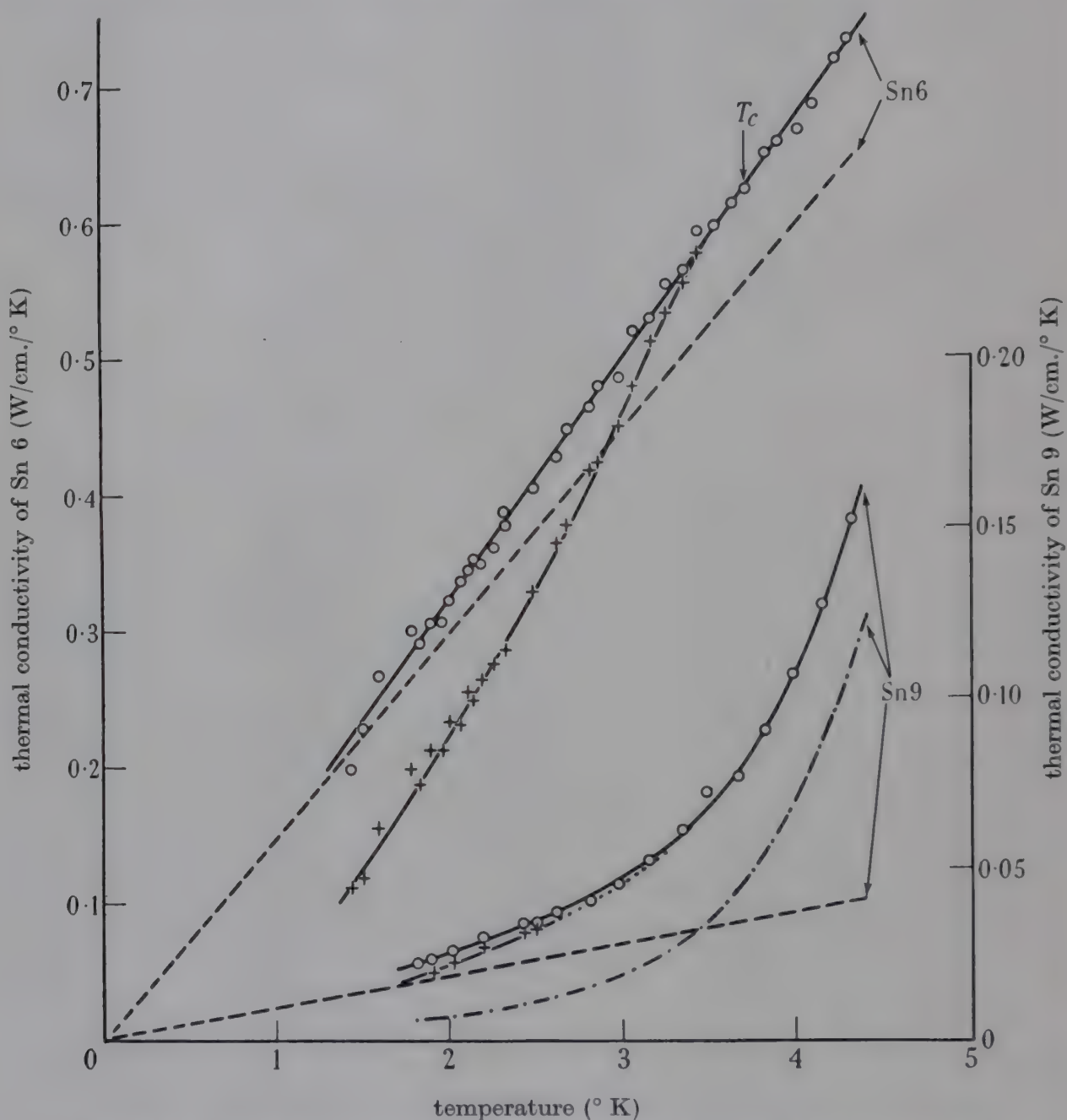


FIGURE 4. Thermal conductivity of tin specimens Sn 6 and Sn 9. $\circ$ normal state, $+$ superconducting state, ---- $L_0 T / \rho_0$, -.-.- lattice conductivity in normal state

above 14° K. These workers observed that as a pure metal is cooled from room temperature, K_n increases and eventually passes through a maximum, the position and height of which depends mainly on the Debye temperature and impurity content of the metal. The present results indicate that spectroscopically pure tin has a maximum at about 4° K, which is moved to higher temperatures, outside the liquid helium range, by the addition of quite small amounts of impurity. The turning point for spectroscopically pure mercury, on the other hand, probably lies below

1° K, and quite large amounts of impurity, such as contained by Hg 6, are necessary to shift the maximum into the liquid helium range.

A question of particular interest is the behaviour of the superconducting curve in the neighbourhood of T_c , which was studied in some detail for the tin and mercury

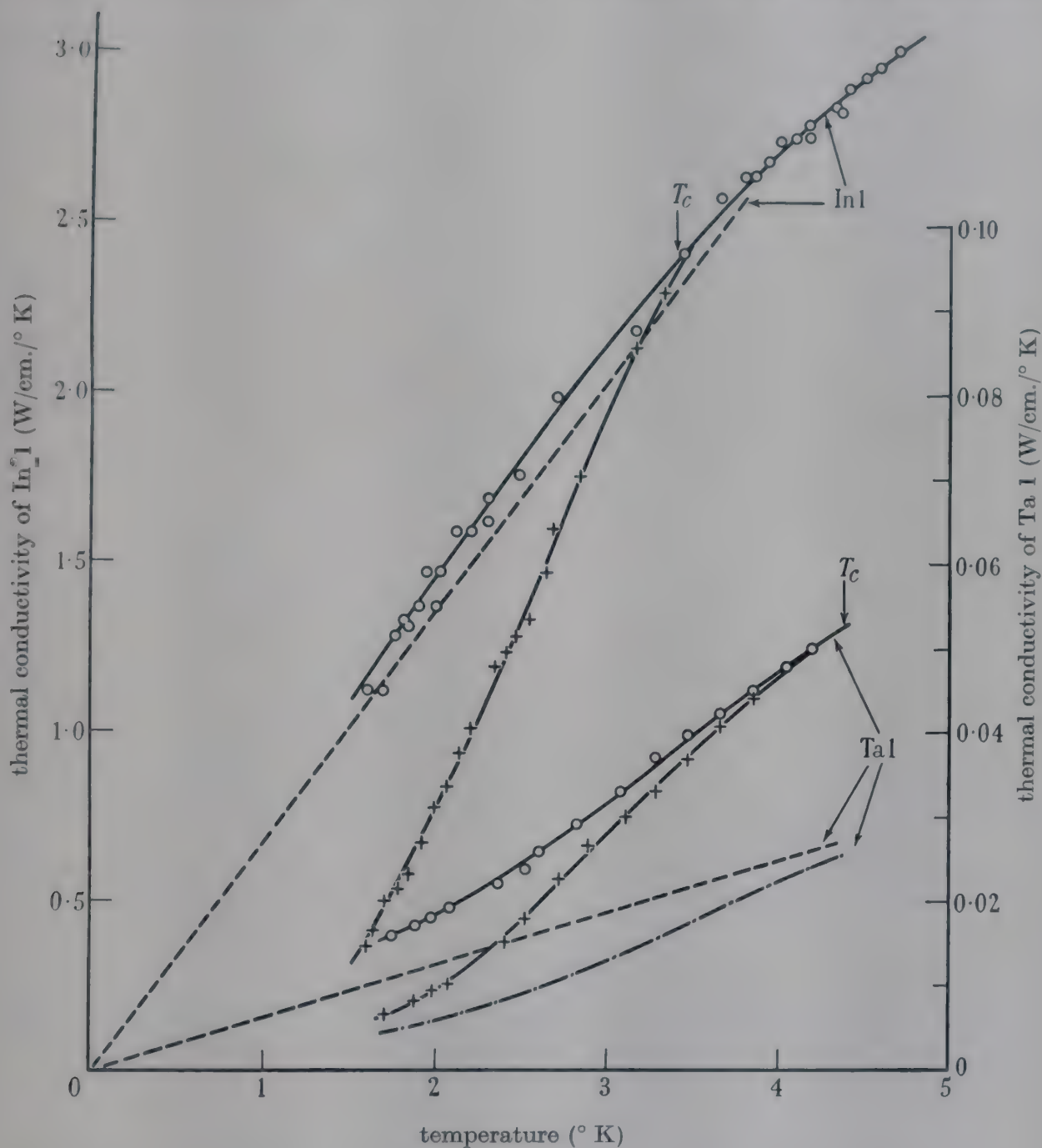


FIGURE 5. Thermal conductivity of indium specimen In 1 and tantalum specimen Ta 1. $\circ$ normal state, $+$ superconducting state, ---- $L_0 T / \rho_0$, -.-.- lattice conductivity in normal state.

specimens. Although changes of K of 0.5 % or less could have been detected, no sudden change was observed for either superconducting or normal specimens on lowering the temperature through the transition point, so that any discontinuity existing at T_c must be an extremely small one. There was definite evidence, however, for a marked discontinuity of the gradient of K for mercury in zero magnetic field

at the transition point, as is shown by the small part of the main curve for Hg 1 enlarged in the right-hand top corner of figure 6. No such discontinuity was observed for any of the tin specimens examined.

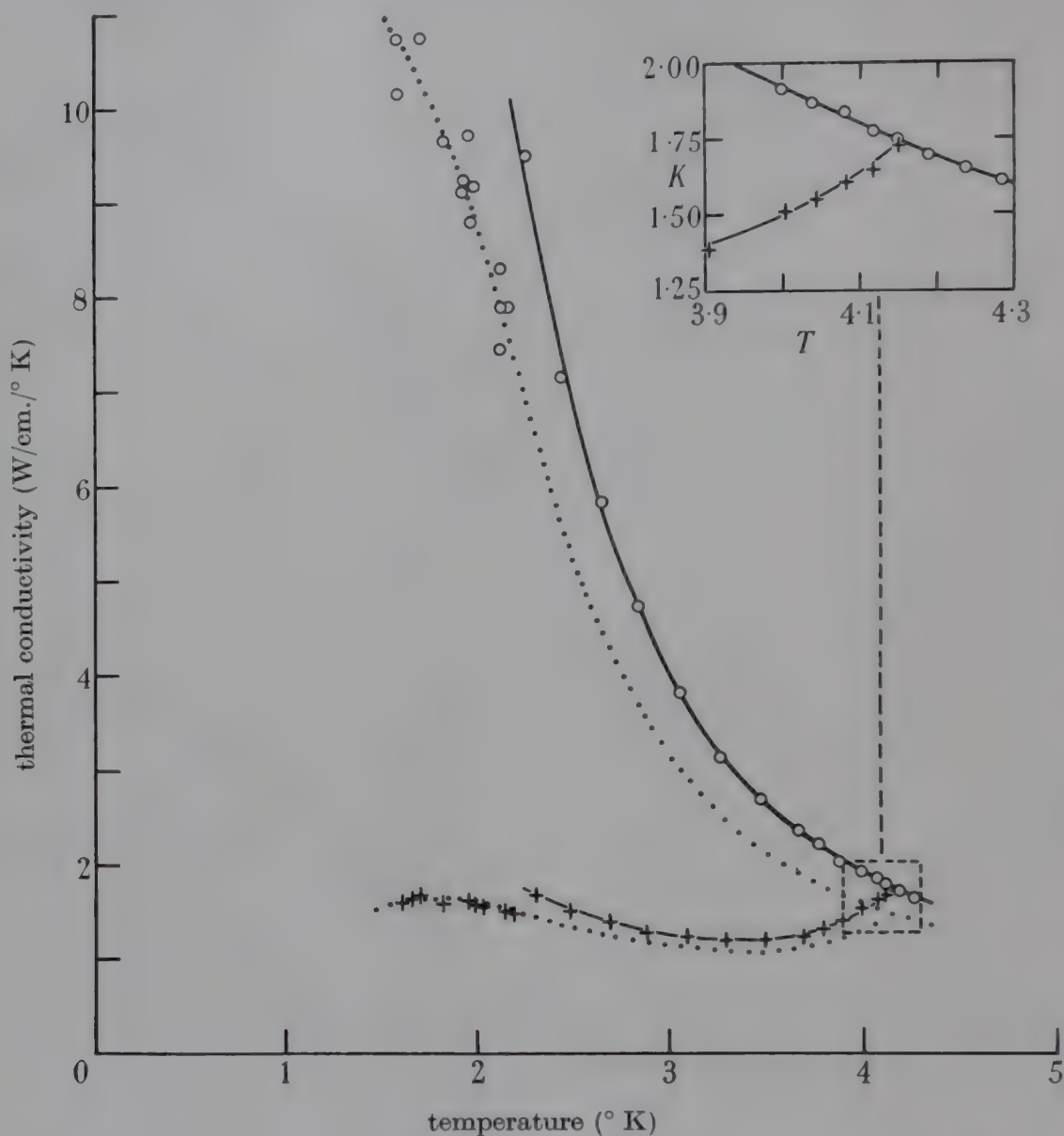


FIGURE 6. Thermal conductivity of mercury specimens Hg 1 (full curves) and Hg 2 (dotted curves and experimental points below 2.2°K). Inset shows detailed behaviour of Hg 1 near to T_0 . $\circ$ normal state, $+$ superconducting state.

4. DISCUSSION OF BEHAVIOUR OF NORMAL METALS

4.1. Outline of theory

Heat-conduction experiments at ordinary temperatures led many years ago to the recognition of two important components in the thermal conductivity of solids, first, the electronic conductivity K_e due to transport of heat by the motion of electrons, which is the dominant mechanism in pure metals, and secondly, the lattice conductivity K_q due to heat transmission by direct interaction between lattice ions, which in alloys may be comparable in importance with electronic conduction and is the sole heat-conduction mechanism in electrical insulators. As a result of theoretical

studies of the temperature variation of K_e and K_g in different types of solid, Peierls (1929), Sommerfeld & Bethe (1933), Wilson (1937) and Makinson (1938) obtained formulae which are in reasonably good agreement with the observed behaviour of thermal conductivity at ordinary temperatures, but which have not been seriously

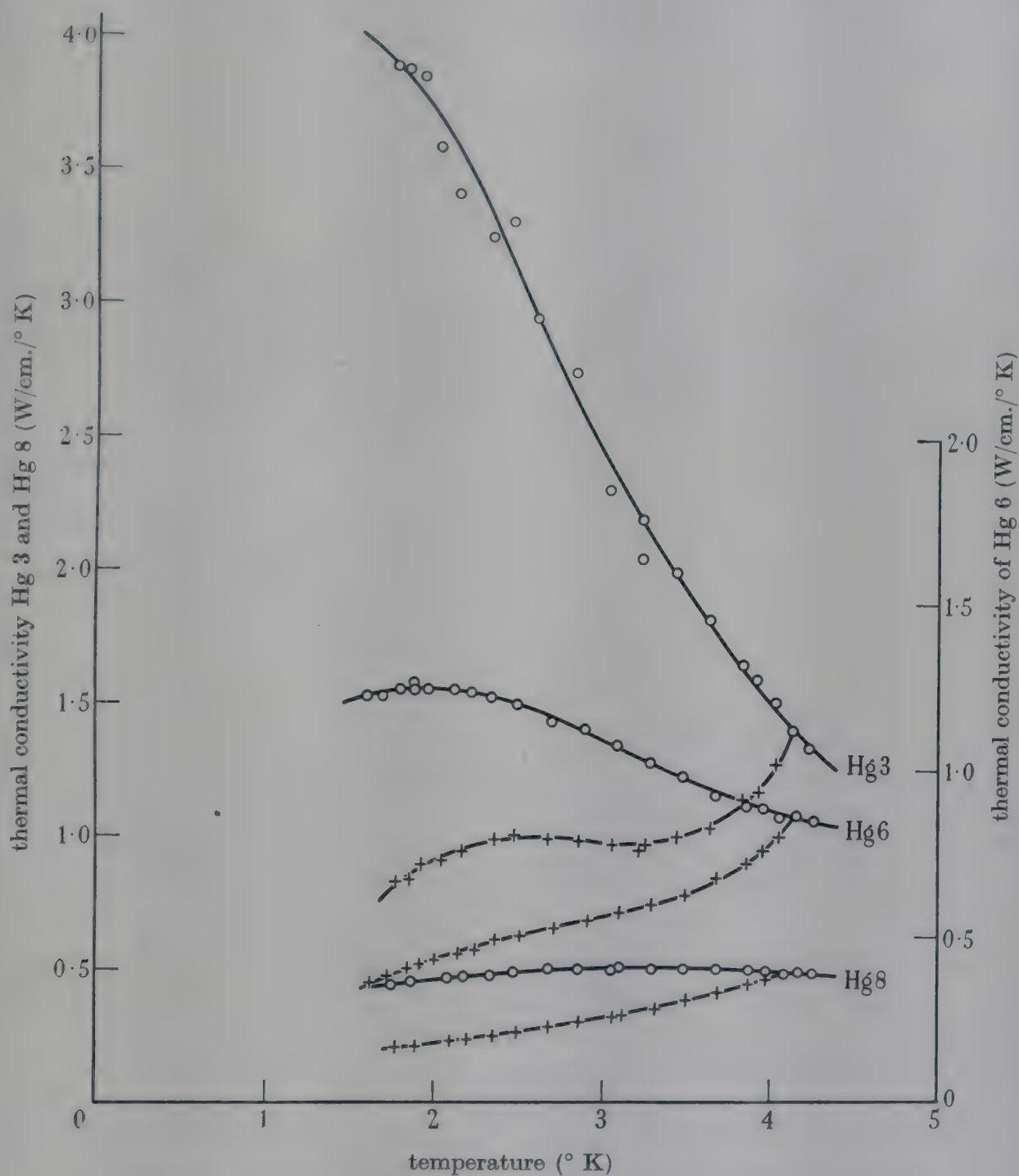


FIGURE 7. Thermal conductivity of mercury specimens Hg 3, Hg 6 and Hg 8.
○ normal state, + superconducting state.

tested at low temperatures owing to the dearth of accurate experimental data. The work of Makinson includes a survey of the existing theoretical results, and therefore forms a convenient starting-point for discussion of data obtained in the present experiments. Makinson points out that the electronic thermal resistance of a normal

metal, $1/K_{en}$, may be expressed as the sum of two terms, which arise respectively from the following scattering processes:

- (1) scattering of electrons by thermal vibrations of the ionic lattice,
- (2) scattering of electrons by impurity atoms and small-scale lattice defects.

His detailed analysis shows that for a free-electron model of a metal the electronic thermal resistance at temperatures less than one-tenth of the Debye temperature Θ takes the form

$$\frac{1}{K_{en}} = \alpha T^2 + \beta/T, \quad (3)$$

where
$$\alpha = \frac{95.3 N_a^{\frac{2}{3}}}{K_{\infty} \Theta^2}, \quad \text{and} \quad \beta = \rho_0/L_0.$$

Here N_a is the effective number of conduction electrons per atom and K_{∞} is the limiting value of the electronic thermal conductivity at high temperatures. For most metals K_{∞} is very nearly the conductivity at room temperature.

Similarly, the lattice thermal resistance, $1/K_{gn}$, may be expressed as the sum of terms arising from the following mechanisms of scattering lattice waves:

- (1) scattering by electrons,
- (2) scattering by crystal boundaries,
- (3) scattering by impurity atoms and small-scale lattice defects,
- (4) mutual scattering of lattice waves, a negligible effect at low temperatures.

If the resistance is largely determined by either of the first two mechanisms, the lattice thermal resistance at temperatures less than $\frac{1}{10}\Theta$ may be written

$$\frac{1}{K_{gn}} = A/T^2 + B/T^3 + CT, \quad (4)$$

the three contributions arising, in order, from the first three scattering mechanisms listed above. The exact form of A , B and C need not concern us here. On the assumption that K_{en} and K_{gn} derived from equations (3) and (4) must be added to give the total normal thermal conductivity, K_n , for a metal at low temperatures, the resulting general expression for the total conductivity has a fairly complex form. There are, however, certain ranges of temperature, composition, crystal size, etc., in which some of the individual terms in (3) or (4) are negligible, so that the other terms may be studied separately. When examined from this point of view, the specimens in table 1 fall naturally into two main groups which will be discussed in detail in the next two sections. The first group comprises those specimens with impurity contents less than about 0.1 % and fairly large crystal grain size, where K_{en} is so large compared to K_{gn} at liquid helium temperatures that the latter may be neglected. The second group includes specimens with impurity contents greater than about 0.1 % for which K_{gn} is comparable to K_{en} .

4.2. *Electronic conduction*

The maximum possible theoretical value of lattice conductivity is determined by the first term on the right-hand side of equation (4), which in contrast to the other terms is dependent only on fundamental constants of the metal. For a typical metal this term gives a value of K_{gn} of about 10^{-2} W/cm./degree at a few degrees absolute,

which is considerably less than most of the experimental values of K_n shown in figures 3 to 7. Evidence will be presented later to show that actual values of lattice conductivity are about 10 times higher than the above theoretical estimate, but even so it appears that K_{gn} is negligible in comparison with K_{en} for the specimens Sn 2, Sn 3, In 1, Hg 1, Hg 2 and Hg 3, which contain less than 0.1 % impurity and have fairly large crystal grains. This is demonstrated by plotting T/K_n against T^3 , as

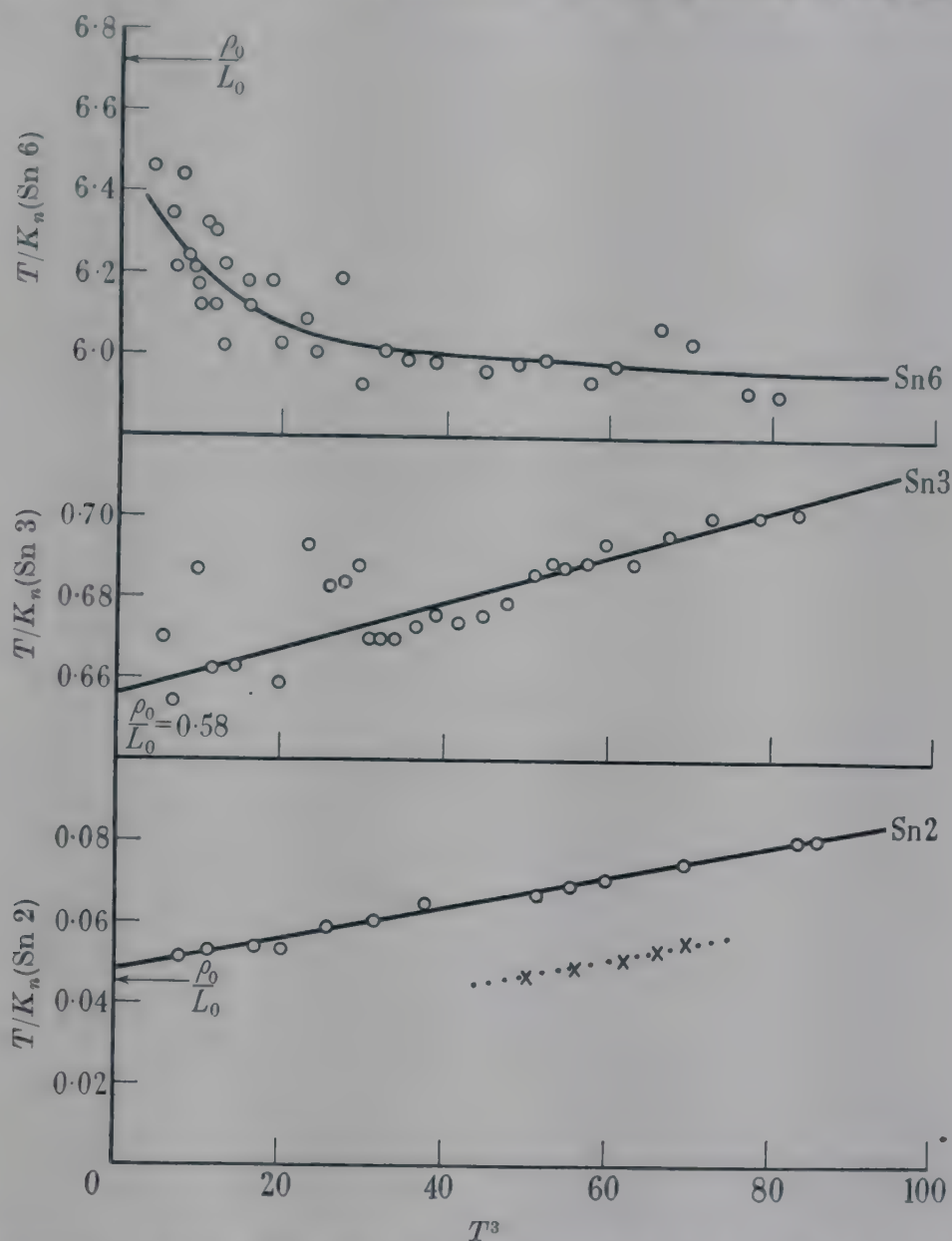


FIGURE 8. Variation of T/K_n with T^3 for tin specimens Sn 2, Sn 3 and Sn 6, and for very pure tin specimen of Rademakers (1949) (dotted curve).

shown for tin and mercury in figures 8 and 9. The experimental points for the above specimens lie on straight lines over most of the range of measurement in agreement with equation (3), so that K_n may be identified with K_{en} . On the other hand, the experimental points for specimens Sn 6, Hg 6 and Hg 8, with impurity contents greater than 0.1 %, show a markedly non-linear behaviour, which may be attributed to the presence of an appreciable component of K_{gn} in addition to K_{en} .

As regards the detailed behaviour of the pure tin specimens, figure 8 shows that the experimental data for Sn 3, Sn 2 and a tin specimen slightly purer than Sn 2

which was investigated by Rademakers (1949), lie on three nearly parallel straight lines with positive intercepts on the T/K_n axis, in spite of the wide difference in

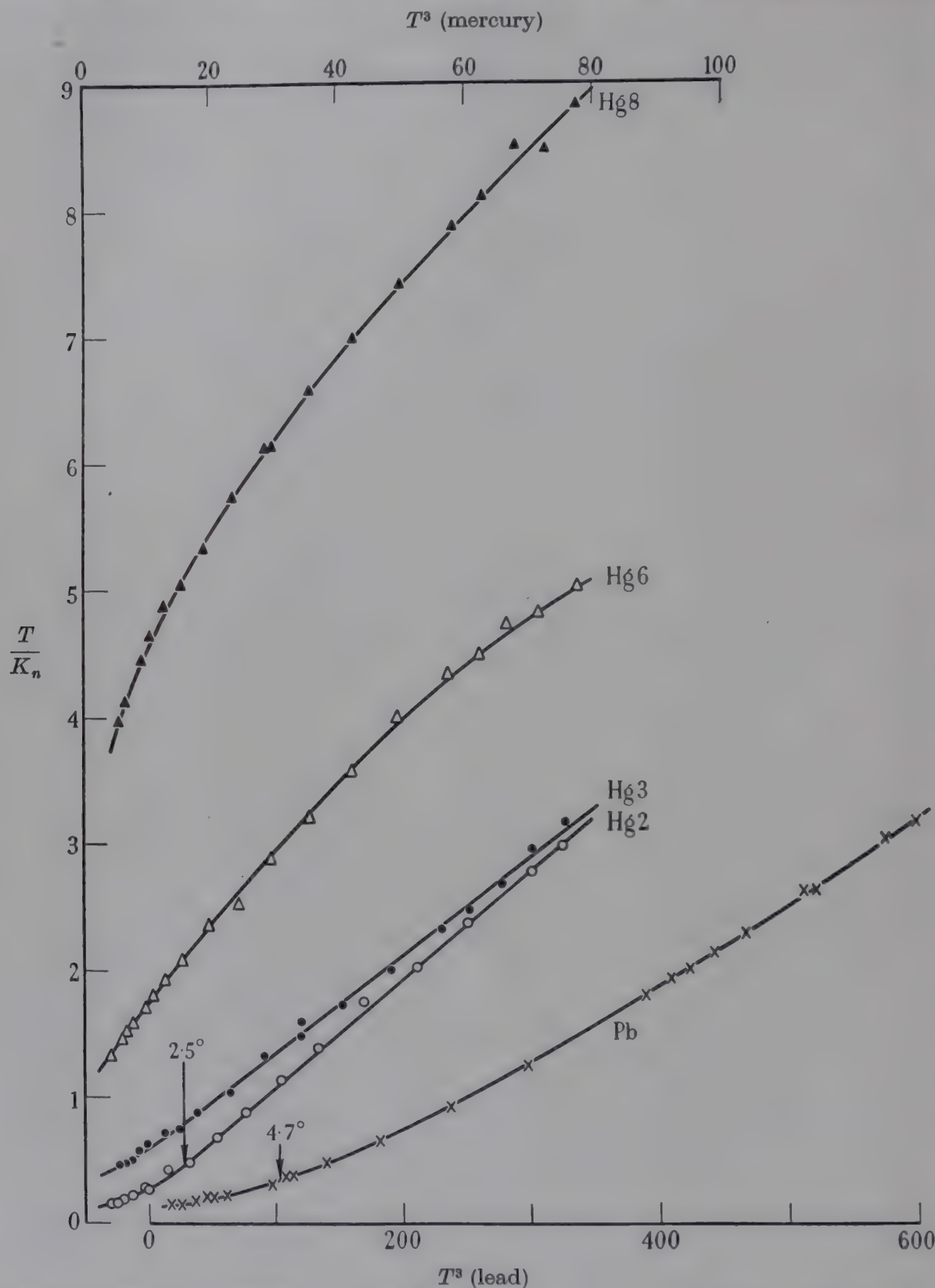


FIGURE 9. Variation of T/K_n with T^3 for mercury specimens. ○ Hg 2, ● Hg 3, Δ Hg 6, ▲ Hg 8, and × for lead specimen of de Haas & Rademakers (1940).

absolute conductivity between Sn 3 and the other two specimens. The linear behaviour of each specimen indicates that the conductivity is essentially electronic with the temperature dependence given by equation (3). The near parallelism of the

straight lines for different specimens shows that the coefficient α is in fact, as theory suggests, a characteristic constant of the metal which is practically independent of impurity content over a wide range of the latter. As a further check on the theory, the experimental value of the intercept on the T/K_n axis, β , may be compared with ρ_0/L_0 determined from the residual electrical resistance. The data in table 1 show that β is several per cent greater than ρ_0/L_0 for Sn 2 and Sn 3, but since the opposite is true for In 1, which also gives a good straight line on a T/K_n against T^3 graph, it seems unlikely that there is any systematic discrepancy.

A close examination of figure 9 shows that the behaviour of the pure mercury specimens is not in as good agreement with theory as that for pure tin, since although T/K_n for Hg 2 and Hg 3 varies linearly with T^3 from the highest temperature of measurement down to about 2.5°K , departure from this linear form occurs at still lower temperatures. It is obvious that a departure of the type observed must occur for Hg 2, owing to the fact that the straight line above 2.5°K has a negative intercept on the T/K_n axis, and cannot therefore represent the behaviour of any actual specimen down to the absolute zero. A negative value of β is quite at variance with theory, since ρ_0/L_0 is essentially a positive quantity. Of interest in this connexion are the data plotted in figure 9 for a spectroscopically pure lead specimen studied by de Haas & Rademakers (1940). The behaviour above 4.7°K is similar to that shown by mercury above 2.5°K , except that the lead curve is markedly non-linear in the upper part of the measuring range. If the experimental points below 4.7°K for lead are plotted on a larger scale than in figure 9 (where many points are omitted to avoid crowding), a straight line is obtained in good agreement with equation (3), and it may be that mercury behaves in a similar manner below 2.5°K , although settlement of this question must await more detailed experimental evidence.

Discrepancies in the detailed application of equation (3) to lead and mercury are not surprising when it is considered that the theory takes no account of the complex electronic structure of these metals, and it is perhaps to be expected that other metals in table 1 will behave anomalously at temperatures outside the range of the present measurements. A similar situation has been exposed by van den Berg (1948) during studies of the electrical resistance of several metals, including lead, at low temperatures. These experiments showed that the electrical resistance due to scattering of electrons by lattice vibrations does not vary with temperature according to the simple power law, cT^5 , predicted by theory. The corresponding case in thermal resistance is a discrepancy between experiment and the T^2 term in equation (3), which is probably the source of the anomalous behaviour of mercury and lead. Both electrical and thermal experiments thus indicate that free-electron theory does not give an adequate detailed representation of the temperature variation of conductivity in actual metals.

As a final test of theory, the experimental values of α for those metals which exhibit a temperature dependence of conductivity in agreement with equation (3) may be used to estimate N_a , the effective number of conduction electrons per atom.

Table 2 contains experimental values of α and estimated values of N_a obtained from the present work and from experiments of previous workers at liquid helium and liquid hydrogen temperatures. Since mercury and lead behave roughly according to

equation (3) at temperatures below 2.5 and 4.7° K respectively, their data are included for the sake of completeness. The copper and tungsten data were obtained by plotting the results of Bremmer & de Haas (1936) on a graph of T/K_n against T^3 . In table 2 are also shown values of N_a derived from the extensive measurements of Gruneisen & Goens (1927) at liquid hydrogen temperatures. Their results were expressed in terms of an empirical formula which at low temperatures takes the form of equation (3), so that by comparison of the coefficients, N_a may immediately be deduced.

TABLE 2

metal	Θ° K	$\alpha \times 10^4$ (experimental)	N_a
Sn	185	3.9	0.013
In	100	19	0.031
Hg	37	200	0.007
Pb*	88	22	0.012
Cu†	320	0.23	0.031
Cu‡	320	—	0.032
W†	346	0.88	0.046
W‡	346	—	0.040
Au†	190	—	0.040
Al†	395	—	0.050
Fe†	470	—	0.092
Pt†	220	—	0.032

* de Haas & Rademakers (1940). † Bremmer & de Haas (1936).
‡ Gruneisen & Goens (1927).

The most striking feature of table 2 is that the values of N_a cluster around 0.03, even for the monovalent metals copper and gold. It is known from other metallic properties that N_a is near unity for the monovalent metals, and is certainly higher than 0.03 for such metals as tin and indium, so that it is unlikely that the values of N_a in table 2 have any physical significance. Of the quantities other than α which are used in the evaluation of N_a , K_∞ is known to a few per cent and Θ is unlikely to be wrong by a factor two in every case, as would be necessary to explain the above discrepancy. It must be concluded, therefore, that in addition to exhibiting anomalies in temperature variation of the type already discussed for mercury and lead, the electronic thermal resistance due to lattice vibration scattering is about 10 times smaller at low temperatures than is predicted by Makinson's theory.

4.3. Lattice conduction

Non-linearity of the T/K_n against T^3 curves for specimens such as Sn 6, Hg 6 and Hg 8, with impurity contents greater than 0.1 %, has already been attributed to the presence of an appreciable lattice component in K_n . Further information on K_{gn} is provided by the specimens Sn 6, Sn 9 and Ta 1, where fairly high values of Θ and ρ_0 together cause the lattice vibration scattering term to be negligible compared to the impurity scattering term in equation (3). This result is unaffected by previously discussed inadequacies of equation (3), since the lattice vibration scattering term is always less than theory predicts. Equation (3) thus becomes

$$K_{en} = L_0 T / \rho_0, \tag{5}$$

so that

$$K_n = L_0 T / \rho_0 + K_{gn}, \tag{6}$$

and K_{gn} may be estimated from the experimental data. The specimens Sn 9 and Ta 1 have lattice components sufficiently large to be estimated reliably with the aid of equation (6), and the resulting curves for K_{gn} are shown in figures 4 and 5. K_{gn} is roughly proportional to T^3 in Sn 9, and to T^2 in Ta 1, but neither specimen shows a simple power-law behaviour over the whole range of measurement. In order to decide which terms are important in equation (4) for the lattice thermal resistance, the specimens must be considered individually.

So far as Sn 9 is concerned, only one of the three terms in equation (4), the first, is likely to be important. The coefficient B determining the resistance due to scattering at crystal boundaries is inversely proportional to the size of the crystallites, which in this specimen was of the order of the dimensions of the rod. This resistive mechanism has been observed in single crystals of quartz (de Haas & Biermasz 1938), for which the resistance was so low as to make it clear that it cannot be the dominant factor here. Again, the steep rise of K_{gn} with increasing temperature seems to rule out the possibility of Sn 9 having an appreciable lattice component due to impurity scattering, since the impurity term on the right-hand side of equation (4) tends to cause K_{gn} to vary inversely as the temperature. A process of elimination thus indicates that scattering of lattice waves by electrons determines the lattice conductivity of Sn 9, even though K_{gn} rises somewhat more rapidly with temperature than the T^2 law given by theory. It is interesting to note that the experimental value of K_{gn} at 4° K is about 10 times greater than the value of the electron scattering term in equation (4) with N_a unity. This discrepancy is of the same order as that found in 4.2 for the reciprocal process, i.e. the scattering of electrons by lattice waves, which suggests that the failure of theory may have the same origin in both cases.

The crystal grain size of Ta 1 is not known, but it was almost certainly smaller than that for Sn 9 and may well have been as small as 10^{-3} cm., which would give a lattice thermal conductivity for boundary scattering of the same order as the observed K_{gn} value. At the same time K_{gn} for Ta 1 is only slightly less than that for Sn 9, so that the effect of electron scattering cannot be ignored. A decision on the relative importance of the two scattering mechanisms will be postponed until the superconducting behaviour has been considered (§5.3).

It should be mentioned that owing to the importance of lattice vibration scattering of electrons for mercury in the liquid helium range of temperatures, equation (3) does not reduce to equation (5) even for large impurity contents, and it is not possible to evaluate K_{gn} with any reasonable accuracy.

5. DISCUSSION OF SUPERCONDUCTING BEHAVIOUR

5.1. *The ratio curves*

The superconducting results for the specimens in table 1 are summarized in figures 10, 11*a* and 12, where the ratio of superconducting to normal conductivity, K_s/K_n , is plotted as a function of reduced temperature, T/T_c . The transition temperatures are obtained from the intersection of the K_s and K_n curves for mercury and from electrical resistance measurements for most of the other specimens. The K_s/K_n data are mainly derived from the smoothed K_s and K_n curves given in

previous figures, but ratios calculated from the experimental values of K_s and K_n are shown for Sn 2 and Sn 3 in order to indicate the typical degree of scatter. It should be mentioned that the Hg 9 curve is probably not as reliable as the others owing to uncertainty in interpretation of the unusual magnetic field curves already described for the specimen; the curve is included to give a qualitative indication of the behaviour of a homogeneous saturated solid solution of indium in mercury.

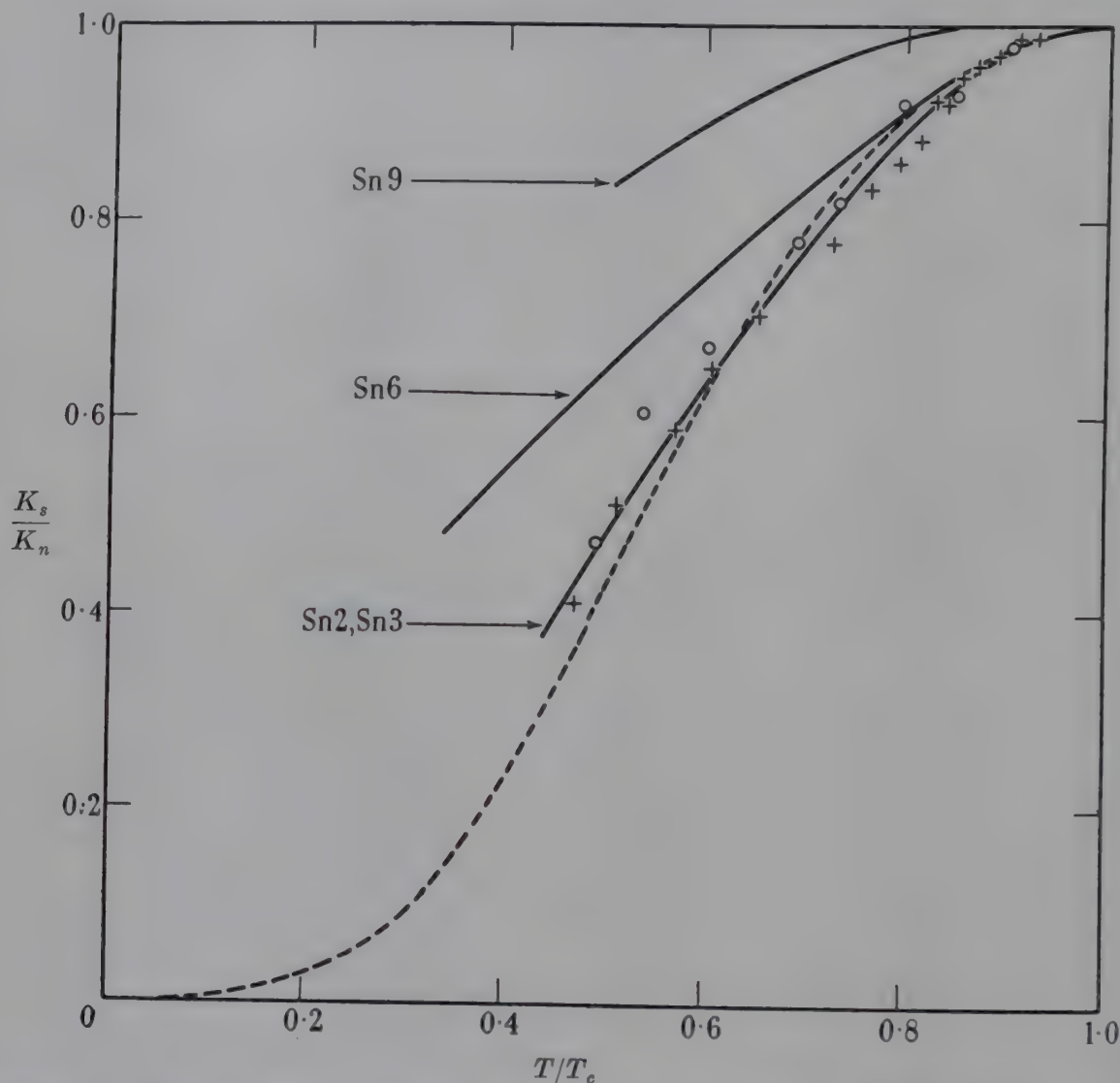


FIGURE 10. Ratio of superconducting to normal conductivity plotted against reduced temperature for tin specimens (points $\circ$, Sn 2, and $+$, Sn 3, show typical scatter). --- Heisenberg (1948).

Detailed theoretical interpretation of the ratio curves on the lines of the previous discussion for normal metals could only be based upon a satisfactory detailed theory of superconductivity, which is not yet available. The following treatment amounts to an empirical classification of the various types of ratio curve, in which it is suggested that for each different mechanism of thermal resistance in a normal metal, that is to say, for each different combination of one of the transport processes with one of its possible scattering processes, there exists a characteristic ratio function peculiar to the particular resistance mechanism and varying with T/T_c in much the same manner for all superconductors; if a particular resistance mechanism is dominant

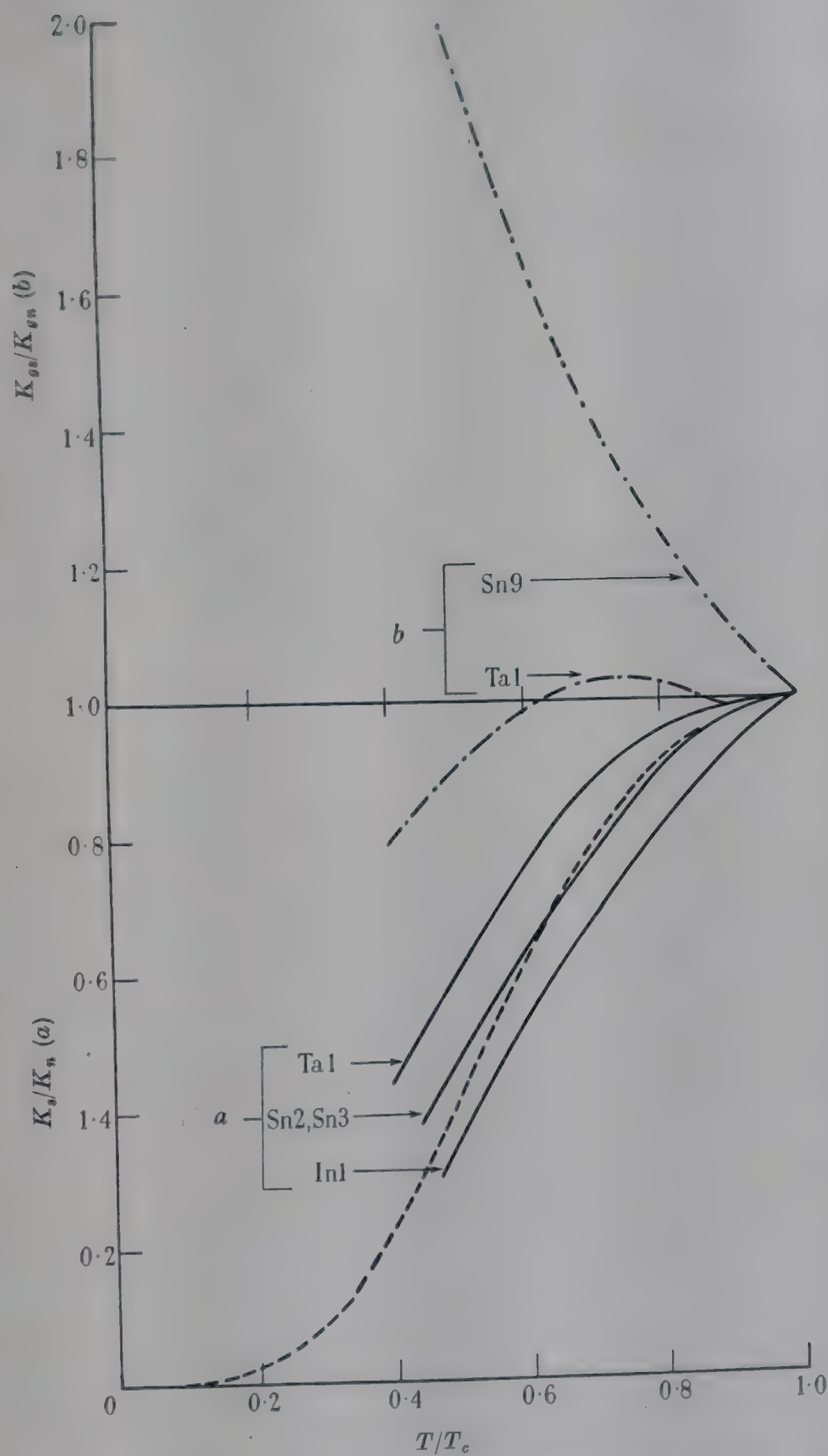


FIGURE 11. *a*, ratio of superconducting to normal conductivity plotted against reduced temperature for Sn 2, Sn 3, In 1 and Ta 1. --- Heisenberg (1948). *b*, ratio of superconducting to normal lattice conductivity plotted against reduced temperature for Sn 9 and Ta 1 (---).

in a normal metal, K_s/K_n has the form of the characteristic ratio function for this mechanism, whereas if several resistance mechanisms are important, K_s/K_n takes a form intermediate between the various characteristic functions. The discussion for

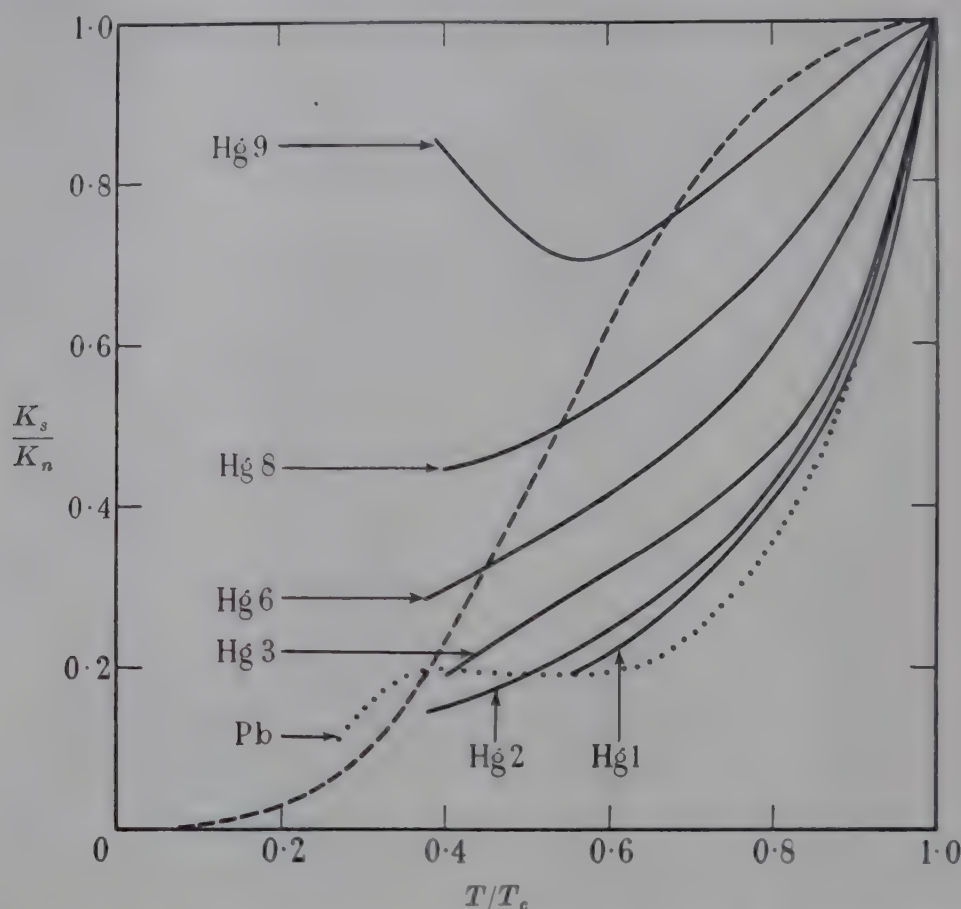


FIGURE 12. Ratio of superconducting to normal conductivity plotted against reduced temperature for mercury specimens (full curves) and lead specimen (dotted curve) of de Haas & Rademakers (1940). --- Heisenberg (1948).

normal metals has indicated four effective thermal resistance mechanisms, which are now listed together with suitable labels for the characteristic ratio functions:

- (1) impurity scattering of electrons, $f(T/T_c)$,
- (2) lattice vibration scattering of electrons, $g(T/T_c)$,
- (3) electron scattering of lattice waves, $h(T/T_c)$,
- (4) crystal boundary scattering of lattice waves, $j(T/T_c)$.

Only meagre experimental data are available for some of these mechanisms, so that certain parts of the above scheme must still be regarded as speculative.

5.2. The ratio f

The most striking feature of figure 10 is that the experimental points for Sn 2 and Sn 3 lie on practically the same curve, in spite of the wide difference in absolute conductivity between the two specimens. This suggests that the ratio may be a characteristic function of temperature, $f(T/T_c)$, under the conditions existing in these specimens, viz. $K_n = K_{en}$ and $\alpha T^2 < \beta/T$. It may be seen from figure 11a that the ratio curve for indium, in which the conditions were similar to those in Sn 2 and Sn 3, is of a similar shape, and it is interesting that there is fair agreement between

these results and the universal theoretical curve proposed by Heisenberg (1948), shown as a broken line in figures 10 and 11*a*. Heisenberg's model of a superconductor is one of a number which depend on the condensation of a proportion of the conduction electrons into a new, superconducting, phase, having the property of passing through the ionic lattice without resistance. On the assumption that the condensed phase plays no part in heat transfer it is clear that the thermal conductivity of a superconductor should be less than that of the normal metal at the same temperature, and this result is brought out of Heisenberg's more detailed calculation.

5.3. The ratio g

A different state of affairs is revealed when the conductivity, though still electronic, is mainly determined by the scattering of electrons by lattice waves, as typified by Hg 1, Hg 2 and Hg 3 and by the lead specimen of de Haas & Rademakers (1940). The ratio curves for these specimens, shown in figure 12, have a general shape quite different from that of $f(T/T_c)$, dropping away roughly as $(T/T_c)^5$ as the temperature is lowered below T_c . If it be assumed that there is a characteristic function $g(T/T_c)$ for the ratio K_{es}/K_{en} in a perfectly pure metal, a simple argument based on equation (3) leads to an expression for the electronic thermal resistance of a superconductor:

$$\frac{1}{K_{es}} = \alpha \frac{T^2}{g(T/T_c)} + \beta \frac{T^{-1}}{f(T/T_c)}. \quad (7)$$

As the temperature is lowered the second term becomes relatively more important, and this type of behaviour may be seen in the experimental curve for lead, which above $0.4T_c$ is of the g -type, and below of the f -type. It will be seen from equation (7) that if β can be made small enough by reduction of impurities it should be possible to observe the g -type of curve, at any rate near T_c , for all metals, even those for which α is low by reason of a high Debye temperature. It is significant, therefore, that data recently published by Rademakers (1949) for a specimen slightly purer than Sn 2 show a marked drop in conductivity just below T_c . Conversely, it should be possible by the addition of impurities to cause a metal such as mercury or lead to exhibit the f -type rather than the g -type of ratio curve. The effect of adding successive amounts of impurities to mercury is shown in figure 12. Clearly the initial shape of the curve tends towards the f -type, but unfortunately so much impurity is needed to accomplish this result that at the lower temperatures the lattice conductivity becomes important, and causes the ratio curve to rise above the f -curve, as will be explained in the next section.

Before leaving this topic it is as well to point out that the fundamental difference between the f and g types of curve is not explained by any theory of superconductivity proposed yet. A discussion of this point is given by Heisenberg (1948).

5.4. The ratios h and j

The K_s/K_n curves for Sn 6, Sn 9 and Ta 1 lie above the curve for Sn 2 and Sn 3 in figures 10 and 11*a*, which is almost certainly due to an appreciable lattice conductivity component in both K_s and K_n for the first three specimens. The normal conductivity of these specimens has already been represented in equation (6) and

by a reasonable modification of this equation, the superconducting conductivity may be written as

$$K_s = f(T/T_c) L_0 T / \rho_0 + K_{gs}, \quad (8)$$

where the first term on the right-hand side is the reduced electronic conductivity derived from the behaviour of Sn 2 and Sn 3. It is unlikely that exactly the same function f applies to both tin and tantalum, but since no data are available on tantalum in the absence of the K_g term, this is the best assumption at present. Using values of K_{gs} estimated from equation (8), and the K_{gn} data already obtained, K_{gs}/K_{gn} is plotted against T/T_c for Sn 9 and Ta 1 in figure 11(b).

Since the lattice waves in Sn 9 are mainly scattered by electrons, the ratio curve for the specimen evidently approximates to the characteristic function $h(T/T_c)$ for the third mechanism of thermal resistance listed in §5.1. The ratio exceeds unity over the whole range of measurement, which finds a qualitative explanation in the electron condensation theory of superconductivity already mentioned. Since the condensation model indicates that the number of electrons which interact with the ionic lattice is smaller in the superconducting state than in the normal state, by the same token the number of lattice wave scattering centres is smaller in the superconducting state than in the normal state, so that K_{gs} exceeds K_{gn} . If this picture is substantially correct, it offers a possible explanation of observations that K_s/K_n is greater than one in certain alloys of lead and in columbium (Bremmer & de Haas 1936; Mendelssohn & Pontius 1937; Mendelssohn & Olsen 1950). Thus if K_{gn} were large compared to K_{en} and the crystal grains were fairly large in these specimens, K_s/K_n would approximate to K_{gs}/K_{gn} , and would probably follow a curve similar to that shown for Sn 9 in figure 11b.

In discussing the K_{gn} results for Ta 1 it was not possible to decide which of the two processes, electron scattering and boundary scattering, was more important for this specimen, but the behaviour of K_{gs}/K_{gn} now suggests that the latter process is dominant. The characteristic ratio $j(T/T_c)$ for lattice conduction with crystal boundary scattering would be expected to be unity on theoretical grounds, since neither the transport nor the scattering mechanism is in this case directly concerned with the electronic state of the metal. The K_{gs}/K_{gn} values for Ta 1 in figure 11b lie in the neighbourhood of one over most of the measuring range, so that the assumption of boundary scattering is the one most consistent with the available evidence.

I should like to express my thanks to Dr D. Shoenberg and Dr A. B. Pippard for suggesting the present problem, and for giving helpful advice throughout the experiments. I am also indebted to Dr E. R. Andrew and Mr R. G. Chambers for measuring the electrical resistance of the specimens. The work was carried out during tenure of a Department of Scientific and Industrial Research grant.

REFERENCES

- Bremmer, H. & de Haas, W. J. 1936 *Physica*, **3**, 672.
 de Haas, W. J. & Engelkes, A. D. 1937 *Physica*, **4**, 325.
 de Haas, W. J. & Biermasz, Th. 1938 *Physica*, **5**, 320.
 de Haas, W. J. & Bremmer, H. 1931 *Commun. Phys. Lab. Univ. Leiden*, 214d, 220b, 220c.
 de Haas, W. J. & Rademakers, A. 1940 *Physica*, **7**, 992.

- Gruneisen, E. & Goens, E. 1927 *Z. Phys.* **44**, 614 and **46**, 151.
 Heisenberg, W. 1948 *Z. Naturforsch.* **3a**, 65.
 Hulm, J. K. 1949 *Nature*, **163**, 368.
 Keesom, W. H. 1932 *Commun. Phys. Lab. Univ. Leiden, Suppl.* 71d.
 Keesom, W. H. & Walstra, W. K. 1940 *Physica*, **7**, 985.
 Keesom, W. H. & Walstra, W. K. 1947 *Physica*, **13**, 225.
 Kistemaker, J. & Keesom, W. H. 1946 *Physica*, **12**, 227.
 Makinson, R. E. B. 1938 *Proc. Camb. Phil. Soc.* **34**, 474.
 Mendelssohn, K. & Olsen, J. L. 1950 *Proc. Phys. Soc. A*, **63**, 2.
 Mendelssohn, K. & Pontius, R. B. 1937 *Phil. Mag.* **24**, 777.
 Peierls, R. 1929 *Ann. Phys., Lpz.*, (5), **3**, 1055.
 Rademakers, A. 1949 *Physica*, **15**, 849.
 Sommerfeld, A. & Bethe, H. 1933 *Handbuch der Physik*, 2nd ed. **24**, pt. 2, p. 545. Berlin: Springer.
 Sondheimer, E. H. & Wilson, A. H. 1947 *Proc. Roy. Soc. A*, **190**, 435.
 Weber, S, Keesom, W. H. & Schmidt, G. 1936 *Commun. Phys. Lab. Univ. Leiden*, 246a.
 Wilson, A. H. 1937 *Proc. Camb. Phil. Soc.* **33**, 371.
 van den Berg, G. J. 1948 *Physica*, **14**, 111.

On the quantum theory of the elementary particles

I. Introduction and classical field dynamics

BY K. V. ROBERTS

*King's College, University of Cambridge, and H. H. Wills Physical
Laboratory, University of Bristol*

(Communicated by N. F. Mott, F.R.S.—Received 14 June 1950)

This paper is the first of a series, concerned with formal aspects of the theory of elementary particles. It introduces a classical approach to the interaction representation of Tomonaga and Schwinger, based upon a generalization of analytical dynamics due to Weiss. General rules are deduced for constructing the interaction Hamiltonian function, and it is explained how the consistency of the representation may be established. The classical part of the theory is developed in detail.

1. INTRODUCTION

One of the current problems in physics is the analysis of the elementary particles. Recent developments in this subject suggest that the 'real' particles investigated in the laboratory must be extremely complicated, since even the properties of electrons can only be obtained approximately, after long and difficult calculation. Although the results of present quantum electrodynamics may not be correct, it would be surprising if those of a future theory were very much simpler, and one may conclude that these particles can only be elementary in a relative sense; something more fundamental must lie behind them. It must be possible to find a theory from which all the properties of elementary particles can be deduced, including, for instance, their mass-ratios and charges, and in which divergences never appear.

To solve this problem, one must find a new set of physical ideas, together with some assumptions upon which a detailed mathematical scheme may be based. Perhaps the new theory will differ considerably from present quantum mechanics, and certainly the solution will not be just a matter of cancelling out divergent expressions from different fields, or finding simple relations between the masses of the elementary particles.

The experimental evidence available at present is limited, and one cannot reasonably expect to reach the truth by pure thought alone. But results can sometimes be obtained by inductive methods, and in theoretical physics these can be of two kinds, physical and mathematical. The second is less powerful, but simpler to use.

There must be some mathematical scheme connected with the new theory which will play a role similar to that of analytical dynamics, or the quantum mechanics of Dirac. We shall try to find out something about this scheme. The results reached so far may perhaps be useful in the immediate problem of eliminating divergences from the theory, and in practical calculations, although they give no clue to the derivation of the elementary constants.

A similar programme forms the basis of the S -matrix theory of Heisenberg, but by itself this theory is rather 'empty', for it contains no principle by which the S -matrix for a particular process can be determined. One cannot expect the S -matrix for all elementary collision processes to be given *a priori*, because these processes, like the particles themselves, are really extremely complicated. In the scheme to which we shall eventually be led the fundamental role is played by another invariant quantity, the Lagrangian function, and from this the S -matrix may be constructed.

At present the quantum theory of fields is the tool which is most commonly used for studying the elementary particles and their interactions. This theory has several advantages over one in which each particle is given a set of co-ordinates and its own separate wave-function. It does not assign a fictitious 'identity' to the particles which is afterwards removed by making the total wave-function symmetrical or anti-symmetrical between them. It avoids difficulties about 'negative energy states' and 'negative probabilities', and allows in a natural way for creation and annihilation processes. Finally, *it treats all particles on a precisely equal footing*, whether they obey Bose or Fermi statistics, are charged or neutral, or have finite or zero mass. Nevertheless, it must be admitted that the field quantities themselves, which are the basic dynamical variables of the theory, have little physical meaning.

The quanta which correspond to these fields have simple properties, and are not the 'real' particles observed in nature, whose properties can only be obtained by renormalization. These real particles are in a sense the result of a balance between the 'ideal' quanta and their environment, the physical vacuum. Freed from all divergence difficulties, this scheme will probably persist in any future theory.

Recently Tomonaga and Schwinger have greatly improved field theory and put it into completely covariant form by introducing the interaction representation. So far, however, no rigorous deduction or mathematical discussion of this representation have been given. For example, wave-functionals are defined on arbitrary, non-parametrized space-like surfaces σ , but it has not been proved that such wave-

functionals exist in general. Also the methods used for deriving the interaction Hamiltonian function have been unnecessarily cumbersome.

In the first two papers of this series a classical approach to this representation will be described, based upon some fundamental ideas of Weiss. The well-known analogies between analytical dynamics and non-relativistic quantum mechanics suggest that the best way to understand the interaction representation might be to develop something similar in classical theory first. The classical field dynamics introduced by Weiss is therefore extended to include canonical transformations and perturbation theory, and then it is possible to prove quite simply that the interaction representation is consistent, and to give a set of rules by which the interaction Hamiltonian function can be written down at once. This approach to the interaction representation is of course rather long, but it is both general and complete, and once the classical field dynamics has been set up it can be applied to any particular theory without further discussion.

It is surprising that the Lagrangian, although fundamental in classical dynamics, has never appeared directly in the quantum theory, for this approach makes the connexion between the two theories so apparent that quantum mechanics might almost be regarded as an integral part of the calculus of variations. In field theory the canonical scheme is very intricate, and the obvious similarity between the properties of the S -matrix, and those of the Lagrangian upon which the theory is based, suggests that one might look for a direct connexion between these quantities. In a later paper of this series it will be proved (using the Heisenberg picture rather than the interaction representation) that the rules for evaluating the Feynman-Dyson S -matrix involve the Lagrangian only, and that the more complicated Hamiltonian function need never be constructed at all. The field Lagrangian may conveniently be divided into two parts:

$$\mathcal{L} = \mathcal{L}^f + \mathcal{L}^i, \quad (1)$$

where $\mathcal{L}^i$, the *interaction Lagrangian*, is the part which depends on the coupling constants. Each kind of vertex which may occur in a graph corresponds to a term in $\mathcal{L}^i$, and conversely. The propagation factors for lines are determined by $\mathcal{L}^f$. 'Surface-dependent terms' and 'contact interactions' have no place in the theory. They appear in the Hamiltonian treatment only to cancel unwanted terms which arise in other ways, and if this treatment were carried through quite consistently would never lead to any physical effects.

One may now choose a different standpoint and try to set up a Lagrangian theory of the elementary particles, in which fields, wave-functionals and Hamiltonian operators do not appear. At present this is equivalent to field theory, but it is simpler to use, and might be developed further. One application is to equivalence theorems for nucleon-meson interactions, which can be proved quite simply by Lagrangian methods.

Finally, although our treatment is closely connected with Feynman's, the word 'Lagrangian' is used in different senses. A classical particle theory is usually deduced from a Lagrangian L ; it may be quantized, and a Schrödinger equation constructed which involves the corresponding particle Hamiltonian H . On the other hand, the

same equation is equivalent to the Lagrangian field equations deduced from the field Lagrangian $\mathcal{L}$. To compare our work with that of Feynman one should compare H with $\mathcal{L}$.

2. A CLASSICAL APPROACH TO THE INTERACTION REPRESENTATION

The interaction representation introduced by Tomonaga and Schwinger (Tomonaga 1946; Koba, Tati & Tomonaga 1947*a, b*; Kanesawa & Tomonaga 1948*a, b*; Schwinger 1948, 1949*a, b*), which has led to results of great importance in quantum electrodynamics, was derived from the non-covariant Heisenberg-Pauli theory. No rigorous mathematical discussion of the representation has yet been given, and the methods employed for finding the interaction Hamiltonian function $\mathcal{H}^i$ have been rather indirect (Kanesawa & Tomonaga 1948*a, b*; Belinfante 1949).

On purely logical grounds it is unsatisfactory that a theory which is completely covariant should be constructed from others which are not, and one ought to look for some way of setting up the interaction representation which is covariant at every stage. It seems to have occurred at about the same time to Matthews (1949), Chang (1949), and the present author that the proper starting-point is the remarkable form of classical field dynamics introduced by Weiss (1936, 1938), which has been undeservedly neglected for many years. Matthews has already given a set of simple rules for constructing $\mathcal{H}^i$. We shall show how the theory may be developed in a general way, from the starting-point of Weiss to the interaction representation of Tomonaga and Schwinger, for an arbitrary set of fields. This classical approach leads naturally to the correct form for $\mathcal{H}^i$, and makes it quite clear that the whole procedure is consistent.

A general introduction to the method is given here, and the third section contains a more detailed treatment of the classical part. The quantum part, together with some classical considerations omitted here to save space, is reserved for part II.

It is well known that non-relativistic quantum mechanics is similar in many ways to the canonical theory of analytical dynamics, a formalism derived from a Lagrange principle, and therefore associated with a problem in the calculus of variations involving several dependent variables q^i , and one independent variable t . The analogy between the classical and quantum theories is really a triple one.

First, the mathematical scheme upon which quantum mechanics is based is closely related to the canonical theory. Secondly, a given classical system can always be taken over into wave mechanics by Schrödinger's method, which corresponds to the transition from geometrical to wave optics. And finally, the quantum dynamics for these systems, and certain others, can be set up by a formal method, due to Heisenberg and Dirac.

This method starts from the canonical formalism of a related classical theory (derived originally from a Lagrange principle) and makes a formal correspondence between the two schemes. Canonical co-ordinates and momenta, the Hamiltonian function, canonical equations of motion and canonical transformations appear in both, while Poisson-bracket relations go over to commutation rules, and the Hamilton-Jacobi equation to that of Schrödinger.

In the field theory of Heisenberg and Pauli the method is extended to a continuous infinity of field functions $\phi^r(x^r)$ ($r = 1, 2, 3$), and a single independent variable t . But although the results are covariant, the method itself is not, since it depends on the choice of a particular set of surfaces $t = \text{const.}$, related to the Lorentz frame.

Now this is really an artificial treatment, since field theory is in fact not derived from a Lagrange principle of the simple kind discussed above, but from a more general one involving a *finite* number of dependent variables $\phi^z(x)$ and *four* independent variables x^μ ($\mu = 0, 1, 2, 3$). For instance, the Lagrange equations are partial differential equations of the second order, not ordinary differential equations as in particle dynamics. The most satisfying approach to the problem would be to develop a canonical formalism based upon a variation principle of this generalized kind, and then see whether an analogous form of quantum mechanics can be constructed. This is the programme initiated by Weiss (1936, 1938). The general kind of quantum mechanics to which it leads, and the interaction representation, are each particular examples of the relativistic quantum mechanics for localizable dynamical systems introduced by Dirac (1948).

A system of interacting classical fields is described by n functions

$$\phi^\alpha(x) \quad (\alpha = 1, 2, \dots, n).$$

α is some tensor or spinor index, and (x) a point of space-time with four co-ordinates x^μ . Upper indices are contravariant and lower indices covariant, and the metrical tensor $g_{\mu\nu}$ has the components $g_{00} = 1$, $g_{11} = g_{22} = g_{33} = -1$, $g_{\mu\nu} = 0$ ($\mu \neq \nu$). One may think of the $\phi^\alpha(x)$ as defining a 4-dimensional surface in an $(n+4)$ -dimensional (ϕ, x) -space. This surface describes the complete history of the system, and is analogous to the 1-dimensional path in the (q, t) -space of particle dynamics. It is an extremal of the variation principle

$$\delta I = \delta \int_D^{(4)} \mathcal{L}(\phi^\alpha(x), \phi_\mu^\alpha(x)) dx = 0 \quad (2)$$

for those variations of the $\phi^\alpha(x)$ which vanish on the boundary of an arbitrary domain D in space-time. The Lagrangian $\mathcal{L}$ is a scalar (density) formed from the $\phi^\alpha(x)$ and their first derivatives, and the integral I is called the action function. $dx = dx^0 dx^1 dx^2 dx^3$, and the index (4) on the integral sign represents the number of integrations to be carried out. $\phi_\mu^\alpha(x) \equiv (\mathbf{d}/\mathbf{d}x^\mu) \phi^\alpha(x)$; the usual notation for partial differentiation is reserved for another process.

The Lagrangian equations which follow from (2) are

$$\frac{\partial \mathcal{L}}{\partial \phi^\alpha} - \frac{\mathbf{d}}{\mathbf{d}x^\mu} \left(\frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha} \right) = 0, \quad (3)$$

which are n partial differential equations of the second order. (Higher-order equations are not discussed here, but the treatment is very similar.) There are several distinct varieties of equation in this general class, all analogues of the ordinary differential equation of the second order met with in particle dynamics. They have different properties, but each is some generalization of a corresponding property of the original equation. For example, either 1-point or 2-point boundary conditions are

appropriate for ordinary second-order linear equations. After generalization the first kind is correct for hyperbolic equations, and the second for elliptic ones.

In dynamics only 1-point conditions have any physical meaning, and it is not surprising that a canonical formalism is only possible for hyperbolic equations. In field theory the $\phi^\alpha(x)$ satisfy the inhomogeneous equations

$$-\left\{\frac{\partial \mathcal{L}^f}{\partial \phi^\alpha} - \frac{\mathfrak{d}}{\mathfrak{d}x^\mu} \left(\frac{\partial \mathcal{L}^f}{\partial \phi_\mu^\alpha} \right)\right\} = \frac{\partial \mathcal{L}^i}{\partial \phi^\alpha} - \frac{\mathfrak{d}}{\mathfrak{d}x^\mu} \left(\frac{\partial \mathcal{L}^i}{\partial \phi_\mu^\alpha} \right), \quad (4)$$

which are linear, since $\mathcal{L}^f$ is quadratic, and hyperbolic because $\mathcal{L}$ is relativistically invariant. To the determination of a solution in particle dynamics, by the values of the co-ordinates and velocities at some instant of time t_0 , corresponds the determination of a solution in field dynamics by the $\phi^\alpha(x)$ and their normal derivatives $\phi'^\alpha(x)$ on some space-like surface σ_0 . Thus $2n$ numbers are replaced by $2n$ functions, and an instant of time by a space-like surface. In order to define these functions it is convenient to set up a 3-dimensional co-ordinate system on each surface, that is, to parametrize it. A parametrized space-like surface $x^\mu = x^\mu(u^1, u^2, u^3)$ will be denoted by S , a non-parametrized surface by σ . Two surfaces S which differ only by a change in their parametrization must nevertheless be regarded as distinct. This parametrization is useful in the general theory, but fortunately does not appear in the interaction representation.

The state of a system may equally well be determined by $2n$ canonical co-ordinate and momentum functions $\phi^\alpha(u)$, $\pi_\alpha(u)$ on some S_0 . The functions $\pi_\alpha(u)$ correspond to the momenta p_i , and are defined by

$$\pi_\alpha = \frac{\partial \mathcal{L}}{\partial \phi'^\alpha} = \frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha} N^\mu, \quad (5)$$

where $\phi'^\alpha = \phi_\mu^\alpha N^\mu$, and N_μ is a normal whose magnitude depends on the parametrization (§ 3, equation (15)). The $\phi^\alpha(u)$ and $\pi_\alpha(u)$ satisfy a set of canonical equations involving four Hamiltonians H , h_r , each of which is a functional of the canonical functions:

$$H = \int_S^{(3)} \mathcal{H}(\phi^\alpha, \tilde{\phi}_r^\alpha, \pi_\alpha) du, \quad h_r = \int_S^{(3)} h_r(\phi^\alpha, \tilde{\phi}_r^\alpha, \pi_\alpha) du. \quad (6)$$

Here $\tilde{\phi}_r^\alpha = (\mathfrak{d}/\mathfrak{d}u^r) \phi^\alpha$, and the Hamiltonian functions $\mathcal{H}$, h_r are defined by

$$\mathcal{H} = \mathcal{L} - \pi_\alpha \phi'^\alpha, \quad h_r = -\pi_\alpha \tilde{\phi}_r^\alpha. \quad (7)$$

The canonical equations are equivalent to the Lagrange equations (3).

All the familiar results of particle dynamics may be extended without difficulty. For instance, the transformation from the functions $\phi^\alpha(u)$, $\pi_\alpha(u)$ on one surface S_1 to the corresponding set on another surface S_2 is canonical, and preserves certain Poisson-bracket relations, of which the most significant are

$$[\pi_\alpha(u), \phi^\beta(u')] = \delta_\alpha^\beta \delta(u - u'). \quad (8)$$

Any other transformation which preserves P.B. relations is canonical, and will provide a new set of canonical equations. The equations of motion can be written in P.B. form, and one may introduce a set of Hamilton-Jacobi equations

$$\frac{\delta W}{\delta w(u)} + \mathcal{H}(u) = 0, \quad \frac{\delta W}{\delta w(u)} + h_r(u) = 0, \quad (9)$$

where W is a functional of the surface S and of the functions $\phi^z(u)$ and $\pi_z(u)$, and w is a locally normal co-ordinate. $\delta/\delta w(u)$ and $\delta/\delta u^r(u)$ stand for functional differentiation with respect to the co-ordinates of a point (u) of S ; the former is independent of the parametrization, and exactly the same as Schwinger's $\delta/\delta\sigma(x)$.

With the help of this scheme two alternative forms of quantum mechanics may be constructed, the Heisenberg and Schrödinger pictures. In the first the state-vector is constant, and the dynamical variables satisfy Heisenberg equations of motion derived from the classical P.B. equations, although since they also satisfy equation (3) it is possible to dispense with most of the canonical formalism. In the second the canonical variables are constant, and the state-vector varies with S according to

$$i\hbar \frac{\delta|\Psi, S\rangle}{\delta w(u)} = \mathcal{H}(u) |\Psi, S\rangle, \quad i\hbar \frac{\delta|\Psi, S\rangle}{\delta u^r(u)} = h_r(u) |\Psi, S\rangle. \tag{10}$$

For equations (9) or (10) to be consistent it is necessary that $\mathcal{H}$ and h_r should satisfy certain integrability conditions given by Dirac (1948), but Hamiltonian functions which are derived originally from a classical Lagrangian will necessarily have this property, simply because they can be regarded as functional derivatives of the action integral I . This ensures the consistency of the theory.

To obtain the interaction representation we introduce a classical perturbation theory, in which the new variables ϕ^{*z} and π_z^* depend on two surfaces S and T . The new canonical variables are equal to the old when $T = S$, and if S is fixed and T allowed to vary they satisfy a set of 'free' canonical equations, based on the functions $\mathcal{H}^f, h_r^f$ of the new variables. Here $\mathcal{H}^f$ and h_r^f are the Hamiltonian functions derived from that part of $\mathcal{L}$ which is independent of the coupling constants, $\mathcal{L}^f$.

If T is fixed, and S allowed to vary, the appropriate Hamiltonian functions, again of the new variables, are $\mathcal{H}^i$ and h_r^i , defined by

$$\mathcal{H}^i = \mathcal{H} - \mathcal{H}^f, \quad h_r^i = h_r - h_r^f. \tag{11}$$

It is, of course, essential that these functions should be expressed in terms of the canonical variables when (11) is carried out.

It is now easy to see that h_r^i is always identically zero, since by (7) h_r and h_r^f are the same. Therefore the new variables do not depend on the parametrization of S , and define a new set of field functions $\phi^{*z}(x; \sigma)$ which obey the free Lagrange equations for variations of x . By solving these equations one can introduce invariant P.B. relations between the field functions at different points of space-time.

We translate this scheme into quantum mechanics by using the Heisenberg picture for variations of T or x , and the Schrödinger picture for variations of S or σ . The state-vector then satisfies

$$i\hbar \frac{\delta|\Psi, \sigma\rangle}{\delta w(u)} = \mathcal{H}^i(u) |\Psi, \sigma\rangle, \tag{12}$$

and does not depend on the parametrization. The variables satisfy the free Lagrange equations and commutation rules and do not depend on σ . We have therefore reached the interaction representation. To prove that (12) is integrable, we need only show that $\mathcal{H}^i$ must satisfy the integrability conditions whenever $(\mathcal{H}, h_r)$ and

$(\mathcal{H}^f, h_r^f)$ do so. It makes no difference whether or not $\mathcal{H}^i(u)$ and $\mathcal{H}^i(u')$ commute for two points u and u' on σ . All the canonical formalism and the parametrization may now be dropped completely, $\mathcal{H}^i$ being expressed in terms of the field functions and their derivatives, and the unit normal n_μ . (For a given point (x) on S , $\mathcal{H}(x)$ never does depend on the parametrization.)

Before discussing the theory in detail, the rules are quoted, for writing down $\mathcal{H}^i$ and other functions which appear in the interaction representation. They follow from the argument above.

For $\mathcal{H}^i$:

- (i) Express $\mathcal{H}$ in terms of the canonical variables $\phi^\alpha, \pi_\alpha, \tilde{\phi}_r^\alpha$, using

$$\mathcal{H} = \mathcal{L} - \phi_\nu^\alpha \frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha} N_\nu N_\mu \quad \text{and} \quad \pi_\alpha = \frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha} N_\mu,$$

where $\mathcal{L}$ is the complete Lagrangian.

- (ii) Split off the part $\mathcal{H}^f$ which is obtained in the same way from the free Lagrangian $\mathcal{L}^f$. The remainder will be $\mathcal{H}^i$, expressed as a function of the canonical variables.

- (iii) Express $\mathcal{H}^i$ in terms of the variables of the interaction representation, using

$$\pi_\alpha = \frac{\partial \mathcal{L}^f}{\partial \phi_\mu^\alpha} N_\mu.$$

For any other function, such as the charge-current vector j_μ :

- (i) Express in terms of the canonical variables, using

$$\pi_\alpha = \frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha} N_\mu.$$

- (ii) Re-express this function of the canonical variables in terms of the variables of the interaction representation, using

$$\pi_\alpha = \frac{\partial \mathcal{L}^f}{\partial \phi_\mu^\alpha} N_\mu.$$

In general, though not always, this alters the form of the function; it would be quite wrong to assume that such functions retained their form in the interaction representation. They may obtain terms which depend on n_μ .

The correct symmetry should be maintained at each stage.

3. CLASSICAL FIELD DYNAMICS

Variation principle

Consider a particular n -dimensional surface in (ϕ, x) -space, whose section in the x -space is a region D , and form the action integral

$$I = \int_D^{(4)} \mathcal{L}(\phi^\alpha(x), \phi_\mu^\alpha(x)) dx, \quad (13)$$

the values $\phi^z(x)$, $\phi_\mu^z(x)$ being determined by the surface. Suppose that each point P of the surface is displaced so that its new 'co-ordinates' in (ϕ, x) -space are

$$\left. \begin{aligned} \bar{\phi}^\alpha(P) &= \phi^\alpha(P) + \epsilon \eta^\alpha(P), \\ \bar{x}^\mu(P) &= x^\mu(P) + \delta x^\mu(P). \end{aligned} \right\} \quad (14)$$

The new action integral $\bar{I}$, which is taken over the displaced region $\bar{D}$, will differ from I by an amount

$$\delta I = \epsilon \int_D^{(4)} \left\{ \frac{\partial \mathcal{L}}{\partial \phi^\alpha} \eta^\alpha + \frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha} \eta_\mu^\alpha \right\} dx + \oint_\Sigma^{(3)} \mathcal{L} N_\mu \delta x^\mu du, \quad (15)$$

where Σ is the closed 3-dimensional boundary of D , parametrized by three co-ordinates u^r ($r = 1, 2, 3$), $du = du^1 du^2 du^3$, and the N_μ are the 3-rowed minors of the Jacobian matrix $(\delta x^\mu / \delta u^r)$. N_μ is not the unit normal to Σ , denoted by n_μ ; its magnitude depends on the parametrization. It is chosen so that for changes of the u^r , $N_\mu du$ is an invariant surface element. N_μ is actually a vector density with weight -1 for transformations of the x^μ and weight 1 for transformations of the u^r . Then both terms in (9) are invariant, since $\mathcal{L}$ is a scalar density.

After integration by parts, (15) becomes

$$\delta I = \epsilon \int_D^{(4)} \left\{ \frac{\partial \mathcal{L}}{\partial \phi^\alpha} - \frac{\delta}{\delta x^\mu} \left(\frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha} \right) \right\} \eta^\alpha dx + \oint_\Sigma^{(3)} (X_\mu \delta x^\mu + \pi_\alpha \delta \phi^\alpha) du, \quad (16)$$

where

$$\delta \phi^\alpha(P) = \epsilon \eta^\alpha(P) + \phi_\mu^\alpha(P) \delta x^\mu(P), \quad (17)$$

and

$$\left. \begin{aligned} \rho_\alpha^\mu &= \frac{\partial \mathcal{L}}{\partial \phi_\mu^\alpha}, & U_\nu^\mu &= \mathcal{L} \delta_\nu^\mu - \rho_\beta^\mu \phi_\nu^\beta, \\ \pi_\alpha &= \rho_\alpha^\mu N_\mu, & X_\nu &= U_\nu^\mu N_\mu. \end{aligned} \right\} \quad (18)$$

U_ν^μ and ρ_α^μ are independent of Σ , while π_α and X_ν depend on Σ through the quantity N_μ .

The Lagrangian equations (3) follow from the principle that $\delta I = 0$ for arbitrary variations (14) which vanish for points P lying on the boundary of the surface. The conservation laws for charge, energy, momentum and angular momentum also follow from invariance properties of I .

Canonical formalism

Restrict the surface Σ , so that it consists of two space-like portions S_1 and S_2 , with a time-like part T joining them. Suppose that T is at infinity; then S_1 and S_2 are two arbitrary parametrized space-like surfaces. Introduce a locally normal co-ordinate w for each S , so that in the neighbourhood of S_1 or S_2 interval is expressed by

$$ds^2 = \gamma^{-1} dw^2 + \gamma_{rs} du^r du^s, \quad (19)$$

where

$$\gamma_{rs} = \frac{\delta x^\mu \delta x_\mu}{\delta u^r \delta u^s} \quad \text{and} \quad \gamma = |\det \gamma^{rs}|. \quad (20)$$

The determinant of the transformation from the x^μ to (w, u^r) is 1. (To make N_μ point towards the future on each surface, we assume that u^r and x^r ($r = 1, 2, 3$) always have the same 'sense'.)

$$\text{Define} \quad \phi'^{\alpha}(u) \equiv \frac{\mathfrak{d}\phi^{\alpha}(x)}{\mathfrak{d}w} \quad (\text{normal derivative}) \quad (21)$$

$$\text{and} \quad \tilde{\phi}_r^{\alpha}(u) \equiv \frac{\mathfrak{d}\phi^{\alpha}(u)}{\mathfrak{d}u^r} \quad (\text{spatial derivatives}). \quad (22)$$

In the canonical formalism the functions $\phi^{\alpha}(u)$, $\phi'^{\alpha}(u)$ are replaced by $\phi^{\alpha}(u)$, $\pi_{\alpha}(u)$, the co-ordinates and momenta. Four Hamiltonian functions are defined by

$$\left. \begin{aligned} \mathcal{H}(u) &= \mathcal{L} - \pi_{\alpha} \phi'^{\alpha} = U_{\nu}^{\mu} N_{\mu} \frac{\mathfrak{d}x^{\nu}}{\mathfrak{d}w} = U_{\nu}^{\mu} n_{\mu} n^{\nu}, \\ \hbar_r(u) &= -\pi_{\alpha} \tilde{\phi}_r^{\alpha} = U_{\nu}^{\mu} N_{\mu} \frac{\mathfrak{d}x^{\nu}}{\mathfrak{d}u^r}. \end{aligned} \right\} \quad (23)$$

The corresponding integrals

$$H = \int_S^{(3)} \mathcal{H}(u) du, \quad h_r = \int_S^{(3)} \hbar_r(u) du, \quad (24)$$

are called the *Hamiltonians*; they clearly depend on the surface, and are not invariant integrals of the kind met before. It is significant that for a fixed space-time point (x) lying on the surface, $\mathcal{H}(x)$ does not depend on the parametrization; and that $\hbar_r$ is a purely formal quantity independent of the Lagrangian.

$\mathcal{H}$ and $\hbar_r$ must be expressed in terms of ϕ^{α} , $\tilde{\phi}_r^{\alpha}$ and π_{α} , so that H and h_r are functionals of the functions $\phi^{\alpha}(u)$ and $\pi_{\alpha}(u)$. Sometimes this transformation to canonical variables cannot be carried out at once—there may be some identical relation between the momenta, as in vector-meson theory. If this happens it is necessary to resort to special methods, for instance, to add auxiliary variables, or to treat only a subset of the ϕ^{α} as independent co-ordinates.

In the second method, *the independent co-ordinates depend explicitly on the surface S* . In vector-meson theory, for example, the normal component of the potential is expressed in terms of the other three. Since the treatment is simpler when the co-ordinates do not depend on S in this way, this case will be discussed separately in the second paper. The results are similar.

Consider a variation of the functions $\phi^{\alpha}(u)$, $\phi'^{\alpha}(u)$ and correspondingly of $\phi^{\alpha}(u)$, $\pi_{\alpha}(u)$ on a given S , but no change in S or its parametrization. Then

$$\delta \mathcal{L}(u) = \frac{\partial \mathcal{L}}{\partial \phi^{\alpha}} \delta \phi^{\alpha}(u) - \frac{\mathfrak{d}}{\mathfrak{d}u^r} \left(\frac{\partial \mathcal{L}}{\partial \tilde{\phi}_r^{\alpha}} \right) \delta \phi^{\alpha}(u) + \frac{\partial \mathcal{L}}{\partial \phi'^{\alpha}} \delta \phi'^{\alpha}(u), \quad (25)$$

$$\delta \mathcal{H}(u) = \delta \mathcal{L}(u) - \delta \phi'^{\alpha}(u) \pi_{\alpha} - \phi'^{\alpha} \delta \pi_{\alpha}(u), \quad (26)$$

$$\delta \hbar_r(u) = \delta \phi(u) \frac{\mathfrak{d}\pi_{\alpha}}{\mathfrak{d}u^r} - \tilde{\phi}_r^{\alpha} \delta \pi_{\alpha}(u). \quad (27)$$

Variations in H and h_r are expressed formally by

$$\int_S^{(3)} \delta \mathcal{H}(u) du = \delta H = \int_S^{(3)} \left(\frac{\delta H}{\delta \phi^{\alpha}(u)} \delta \phi^{\alpha}(u) + \frac{\delta H}{\delta \pi_{\alpha}(u)} \delta \pi_{\alpha}(u) \right) du, \quad (28)$$

$$\int_S^{(3)} \delta \hbar_r(u) du = \delta h_r = \int_S^{(3)} \left(\frac{\delta h_r}{\delta \phi^{\alpha}(u)} \delta \phi^{\alpha}(u) + \frac{\delta h_r}{\delta \pi_{\alpha}(u)} \delta \pi_{\alpha}(u) \right) du. \quad (29)$$

From the Lagrangian equations

$$\frac{\delta \pi_\alpha}{\delta w} = \frac{\partial \mathcal{L}}{\partial \phi^\alpha} - \frac{\delta}{\delta u^r} \left(\frac{\partial \mathcal{L}}{\partial \dot{\phi}_r^\alpha} \right). \quad (30)$$

If (30) is used in (25), and the resulting expression for $\delta \mathcal{L}(u)$ substituted in (26), then

$$\delta \mathcal{H}(u) = \frac{\delta \pi_\alpha}{\delta w} \delta \phi^\alpha(u) - \frac{\delta \phi^\alpha}{\delta w} \delta \pi_\alpha(u). \quad (31)$$

Finally, comparing (28), (29) with (31), (27), we obtain the canonical equations

$$\frac{\delta \phi^\alpha(u)}{\delta w} = -\frac{\delta H}{\delta \pi_\alpha(u)}, \quad \frac{\delta \pi_\alpha(u)}{\delta w} = \frac{\delta H}{\delta \phi^\alpha(u)}, \quad (32)$$

$$\frac{\delta \phi^\alpha(u)}{\delta u^r} = -\frac{\delta h_r}{\delta \pi_\alpha(u)}, \quad \frac{\delta \pi_\alpha(u)}{\delta u^r} = \frac{\delta h_r}{\delta \phi^\alpha(u)}, \quad (33)$$

which are equivalent to $2n$ partial differential equations of the first order.

We consider briefly the meaning of (32) and (33). S is a parametrized surface, and so any set of functions $\phi^\alpha(x)$ in space-time—any n -dimensional surface in (ϕ, x) -space—determines $2n$ functions $\phi^\alpha(u)$, $\pi_\alpha(u)$ for each S .

To (q, p) -space, or phase space, corresponds the space P of all sets of $2n$ functions $(\phi^\alpha(u), \pi_\alpha(u))$ of three variables u ; and to the t -axis, the space Ω of all parametrized space-like surfaces S ; $x^\mu = x^\mu(u^r)$.

The $\phi^\alpha(x)$ define a mapping of Ω into P —the analogue of a trajectory in (q, p, t) -space in particle dynamics.

If the parametrization of S alone is altered, the co-ordinates (u) of each space-time point (x) on the surface change and therefore the functions $\phi^\alpha(u)$, $\pi_\alpha(u)$ transform; the new functions $\phi^{*\alpha}(u)$, $\pi_\alpha^*(u)$ are related to the old by

$$\phi^\alpha(u) = \phi^{*\alpha}(u^*), \quad \gamma^{-1} \pi_\alpha(u) = \gamma^{*-1} \pi_\alpha^*(u^*),$$

where (u) , (u^*) are the old and the new co-ordinates respectively of the same space-time point.

If S undergoes a bodily change, $\phi^\alpha(u)$ and $\pi_\alpha(u)$ transform in a more significant manner. Each of these examples is a *canonical transformation*, or mapping of P into itself, satisfying a condition to be defined later.

(32) and (33) define *infinitesimal* canonical transformations. Since $\phi^\alpha(u)$ and $\pi_\alpha(u)$ are functionals of S (of a rather specialized form), one might consider that the transformation defined by the canonical equations should be that induced by the most general infinitesimal deformation of S . Then these equations would involve *functional* derivatives.

The canonical equations may, however, be interpreted in another way. If S is kept fixed, (32) and (33) define the variation of the functions $\phi^\alpha(u)$, $\pi_\alpha(u)$, and so of the $\phi^\alpha(x)$, in the neighbourhood of a point on S . For this purpose it is convenient to have them in their present form. We shall, however, meet canonical variables which are more general functionals of S , and for these it is appropriate to use functional differentiation.

Equations (33) are purely formal, and independent of the particular system under discussion.

Poisson brackets

Extremals of the variation principle (2) may be determined in two ways, either by the Lagrangian equations (3), or by a *boundary formula*. If we assume that the equations (3) are satisfied, and employ the surface Σ defined above, (16) becomes

$$\delta I = \left| \int_{S_1}^{(3)} (X_\mu(u) \delta x^\mu(u) + \pi_\alpha(u) \delta \phi^\alpha(u)) du \right|_{S_1}^{S_2}, \quad (34)$$

$$\text{or} \quad = \left| \int_{S_1}^{(3)} (\mathcal{H}(u) \delta w(u) + h_r(u) \delta u^r(u) + \pi_\alpha(u) \delta \phi^\alpha(u)) du \right|_{S_1}^{S_2}. \quad (35)$$

(The sign of the contribution from the earlier surface S_1 has been altered.)

I may be thought of as a *functional* of the parametrized surfaces S_1 and S_2 , and of a suitably chosen set of $2n$ functions, e.g. $\phi^\alpha(u, S_1)$ and $\phi^\alpha(u, S_2)$; $(\mathcal{H}, h_r)$ or X_μ , and $\pi_\alpha(u, S_1)$, $\pi_\alpha(u, S_2)$ are the corresponding functional derivatives. That is, if a point (u) on S_2 is displaced through a distance $\delta x^\mu = (\delta w, \delta u^r)$,

$$\left. \begin{aligned} \delta I &\equiv \frac{\delta I}{\delta x^\mu(u)} \delta x^\mu(u) = X_\mu(u) \delta x^\mu(u), \\ \text{or} \quad &\equiv \frac{\delta I}{\delta w(u)} \delta w(u) + \frac{\delta I}{\delta u^r(u)} \delta u^r(u) = \mathcal{H}(u) \delta w(u) + h_r(u) \delta u^r(u), \end{aligned} \right\} \quad (36)$$

while for a variation $\delta \phi^\alpha(u)$ at (u) on S_2 ,

$$\delta I \equiv \frac{\delta I}{\delta \phi^\alpha(u)} \delta \phi^\alpha(u) = \pi_\alpha(u) \delta \phi^\alpha(u). \quad (37)$$

Similar results, with a negative sign, hold for variations on S_1 . The equations (36) are to be interpreted as meaning that, for example,

$$\delta I = \int_{S_1}^{(3)} X_\mu(u) \delta x^\mu(u) du \quad (38)$$

for all those infinitesimal deformations of the surface S_2 :

$$S_2: x^\mu(u) \rightarrow x^\mu(u) + \delta x^\mu(u),$$

under which it remains everywhere space-like. This interpretation of functional derivatives with respect to the co-ordinates of a point (u) of S is to be understood throughout.

All the properties of the dynamical system may be deduced from the boundary formula (34) or (35). We shall now introduce the concept of the Poisson bracket, starting from the relative integral invariant. We select two arbitrary, fixed, space-like surfaces S_1 and S_2 , a parameter λ which runs from 0 to 1, and a 1-parameter 'closed' set of $2n$ functions $\phi^\alpha(u, \lambda)$, $\pi_\alpha(u, \lambda)$ on S_1 . By 'closed' we mean that

$$\phi^\alpha(u, 0) = \phi^\alpha(u, 1),$$

$$\pi_\alpha(u, 0) = \pi_\alpha(u, 1).$$

The analogue of this is a closed curve in phase space, at a given time t_1 , which determines a 'tube' of extremals passing through the curve.

I is now a function of λ . We put $\delta\phi^\alpha = \frac{d\phi^\alpha}{d\lambda} \delta\lambda$ in (35), and integrate from 0 to 1. Then $I(0) = I(1)$, and so

$$\int_0^1 d\lambda \int_{S_1}^{(3)} \left(\pi_\alpha \frac{d\phi^\alpha}{d\lambda} \right) du = \int_0^1 d\lambda \int_{S_2}^{(3)} \left(\pi_\alpha \frac{d\phi^\alpha}{d\lambda} \right) du. \quad (39)$$

For a given surface S_1 , S_2 is of course quite arbitrary, and so

$$\int_0^1 d\lambda \int_S^{(3)} \left(\pi_\alpha \frac{d\phi^\alpha}{d\lambda} \right) du \quad (40)$$

is a constant of the motion. It is called the *relative integral invariant*, since it is an integral over a closed region.

We now choose any 2-parameter 'open' family of extremals, with the 1-parameter family above as its boundary, and label it by two parameters (ω_1, ω_2) . By Stokes's theorem, (40) may be transformed to

$$\int^{(2)} d\omega_1 d\omega_2 \int_S^{(3)} \left(\frac{\partial \pi_\alpha(u)}{\partial \omega_1} \frac{\partial \phi^\alpha(u)}{\partial \omega_2} - \frac{\partial \pi_\alpha(u)}{\partial \omega_2} \frac{\partial \phi^\alpha(u)}{\partial \omega_1} \right) du, \quad (41)$$

and this expression is called the *absolute integral invariant*. Since the domain of integration in (ω_1, ω_2) -space is arbitrary, the integrand of (41) must itself be a constant of the motion. It is called the *Lagrange bracket* of ω_1 and ω_2 :

$$\{\omega_1, \omega_2\} = \int_S^{(3)} \left(\frac{\partial \pi_\alpha(u)}{\partial \omega_1} \frac{\partial \phi^\alpha(u)}{\partial \omega_2} - \frac{\partial \pi_\alpha(u)}{\partial \omega_2} \frac{\partial \phi^\alpha(u)}{\partial \omega_1} \right) du. \quad (42)$$

Previously, $\phi^\alpha(u)$, $\pi_\alpha(u)$ were treated as functions of ω_1 and ω_2 , but we may invert the relationship, and think of ω_1 and ω_2 as *functionals* of the $2n$ functions $\phi^\alpha(u)$, $\pi_\alpha(u)$ on S_1 . Then

$$\delta\omega = \int_{S_1}^{(3)} \left(\frac{\delta\omega}{\delta\phi^\alpha(u)} \delta\phi^\alpha(u) + \frac{\delta\omega}{\delta\pi^\alpha(u)} \delta\pi^\alpha(u) \right) du. \quad (43)$$

ω_1 and ω_2 could be any independent quantities connected with the motion, e.g. ω_1 might be $\phi^\alpha(x)$ at some point P of space-time.

The *Poisson bracket* of ω_1 and ω_2 is defined by

$$[\omega_1, \omega_2] = \int_{S_1}^{(3)} \left(\frac{\delta\omega_1}{\delta\pi_\alpha(u)} \frac{\delta\omega_2}{\delta\phi^\alpha(u)} - \frac{\delta\omega_1}{\delta\phi^\alpha(u)} \frac{\delta\omega_2}{\delta\pi_\alpha(u)} \right) du, \quad (44)$$

and is also a constant of the motion. This may be proved by showing that

$$\int_{S_1}^{(3)} \{ \omega_a(u), \omega_b(u') \} [\omega_a(u), \omega_c(u'')] du = \delta_c^b \delta(u' - u''), \quad (45)$$

where the functions $\omega_a(u)$ ($a = 1, \dots, 2n$) form a complete independent set (sufficient, like the $\phi^\alpha(u)$, $\pi_\alpha(u)$ to define an extremal), and we sum over a . The proof is straightforward; we merely insert the definitions (42), (44) and evaluate (Weiss 1938, pp. 115–116). We omit it here to save space.

From (44), it is clear that

$$\left. \begin{aligned} [\pi_\alpha(u), \pi_\beta(u')] &= 0, \\ [\phi^\alpha(u), \phi^\beta(u')] &= 0, \\ [\pi_\alpha(u), \phi^\beta(u')] &= \delta_\alpha^\beta \delta(u - u'). \end{aligned} \right\} \quad (46)$$

Differentiation of the second and third of these equations yields

$$\left. \begin{aligned} [\phi^\alpha(u), \dot{\phi}^\beta_r(u')] &= 0, \\ [\pi_\alpha(u), \dot{\phi}^\beta_r(u')] &= -\delta_\alpha^\beta \delta'_r(u - u'), \end{aligned} \right\} \quad (47)$$

where

$$\delta'_r(u - u') = (\partial/\partial u^r) \delta(u - u').$$

Poisson-bracket equations of motion

If F is any functional of the $\phi^\alpha(u)$, $\pi_\alpha(u)$, or function of the $\phi^\alpha(u)$, $\dot{\phi}^\alpha_r(u)$, $\pi_\alpha(u)$, which does not depend explicitly on the surface,

$$\begin{aligned} \frac{\delta F}{\delta w} &= \int_S^{(3)} \left(\frac{\delta F}{\delta \pi_\alpha(u)} \frac{\delta \pi_\alpha(u)}{\delta w} + \frac{\delta F}{\delta \phi^\alpha(u)} \frac{\delta \phi^\alpha(u)}{\delta w} \right) du \\ &= \int_S^{(3)} \left(\frac{\delta F}{\delta \pi_\alpha(u)} \frac{\delta H}{\delta \phi^\alpha(u)} - \frac{\delta F}{\delta \phi^\alpha(u)} \frac{\delta H}{\delta \pi_\alpha(u)} \right) du, \end{aligned}$$

i.e.

$$\frac{\delta F}{\delta w} = [F, H]. \quad (48)$$

Similarly,

$$\frac{\delta F}{\delta u^r} = [F, h_r]. \quad (49)$$

Canonical transformations

Consider a non-singular mapping of the space P into itself

$$(\phi^\alpha(u), \pi_\alpha(u)) \rightarrow (\phi^{*\alpha}(u), \pi_\alpha^*(u)). \quad (50)$$

The new quantities will be non-singular functionals of the old, and any variation $(\delta\phi^\alpha(u), \delta\pi_\alpha(u))$ of the first set will induce a variation $(\delta\phi^{*\alpha}(u), \delta\pi_\alpha^*(u))$ of the second.

A transformation (50) is called *canonical* if

$$\int_S^{(3)} (\pi_\alpha(u) \delta\phi^\alpha(u) - \pi_\alpha^*(u) \delta\phi^{*\alpha}(u)) du = -\delta W, \quad (51)$$

where δW is the perfect functional differential of a functional W of $\phi^\alpha(u)$, $\pi_\alpha(u)$ and possibly S .

For a set of functions to be canonical, it is not really necessary that they be functions of the u^r . For instance, the Fourier transforms of $\phi^\alpha(u)$, $\pi_\alpha(u)$ do not satisfy this condition, but they do satisfy canonical equations, and are functionals of $\phi^\alpha(u)$, $\pi_\alpha(u)$. (51) might therefore be extended to

$$\int_S p_\beta(y) \delta q^\beta(y) dy - \int_S P_\gamma(z) \delta Q^\gamma(z) dz = -\delta W, \quad (52)$$

where (β, y) , (γ, z) are appropriate discrete and continuous variables.

The following theorems are given without proof.

THEOREM I. *A transformation is canonical if and only if it preserves Lagrange brackets.*

THEOREM II. *A transformation is canonical if and only if it preserves Poisson brackets.*

THEOREM III. The transformation (50) is defined by certain equations which express $\phi^{*\alpha}(u), \pi_\alpha^*(u)$ in terms of $\phi^\alpha(u), \pi_\alpha(u)$.

(a) Suppose that these cannot be solved to give any functional relations between the $\phi^\alpha(u)$ and $\phi^{*\alpha}(u)$ alone, which do not involve the $\pi_\alpha(u)$ or $\pi_\alpha^*(u)$. Express W as a functional of the $\phi^\alpha(u)$ and $\phi^{*\alpha}(u)$.

Then if the transformation is canonical,

$$\pi_\alpha^*(u) = \frac{\delta W}{\delta \phi^{*\alpha}(u)}, \quad \pi_\alpha(u) = -\frac{\delta W}{\delta \phi^\alpha(u)}. \quad (53)$$

Conversely, if W is a functional of the $\phi^\alpha(u)$ and $\phi^{*\alpha}(u)$, and if the $\phi^{*\alpha}(u), \pi_\alpha^*(u)$ are defined in terms of the $\phi^\alpha(u), \pi_\alpha(u)$ by the equations (53), then the transformation is canonical.

(b) Alternatively, it may be possible to solve the defining equations to give a number of relations of the kind

$$\Omega^i[\phi^\alpha(u), \phi^{*\alpha}(u)] = 0 \quad (i = 1, 2, \dots, r, \text{ say}). \quad (54)$$

Then if the transformation is canonical,

$$\left. \begin{aligned} \pi_\alpha^*(u) &= \frac{\delta W}{\delta \phi^{*\alpha}(u)} + \lambda_i \frac{\delta \Omega^i}{\delta \phi^{*\alpha}(u)}, \\ \pi_\alpha(u) &= -\frac{\delta W}{\delta \phi^\alpha(u)} - \lambda_i \frac{\delta \Omega^i}{\delta \phi^\alpha(u)}, \end{aligned} \right\} \quad (55)$$

where the λ_i are functionals of the $\phi^\alpha(u)$ and $\phi^{*\alpha}(u)$. Conversely, if W and the Ω^i are functionals of the $\phi^\alpha(u)$ and $\phi^{*\alpha}(u)$, and if the $\phi^{*\alpha}(u), \pi_\alpha^*(u)$ are defined in terms of the $\phi^\alpha(u), \pi_\alpha(u)$ by the $(2n+r)$ equations (54), (55), (the $r \lambda_i$ being eliminated), then the transformation is canonical.

Infinitesimal canonical transformations

Consider a canonical transformation in which the $\phi^{*\alpha}(u), \pi_\alpha^*(u)$ differ from the $\phi^\alpha(u), \pi_\alpha(u)$, by quantities which are infinitesimal. Then

$$\left. \begin{aligned} \phi^{*\alpha}(u) &= \phi^\alpha(u) + \delta \phi^\alpha(u) = \phi^\alpha(u) + \epsilon \chi^\alpha(u), \\ \pi_\alpha^*(u) &= \pi_\alpha(u) + \delta \pi_\alpha(u) = \pi_\alpha(u) + \epsilon \psi_\alpha(u), \end{aligned} \right\} \quad (56)$$

where ϵ is an arbitrary infinitesimal constant, and $\chi^\alpha(u), \psi_\alpha(u)$ are functions defined on S_1 satisfying conditions to be determined. Since the transformation is canonical,

$$\int_S^{(3)} (\pi_\alpha(u) \delta \phi^\alpha(u) - \pi_\alpha^*(u) \delta \phi^{*\alpha}(u)) du = -\delta W,$$

where W is a functional of the $\phi^\alpha(u)$ and $\pi_\alpha(u)$. After substitution,

$$\epsilon \int_S^{(3)} (\pi_\alpha(u) \delta \chi^\alpha(u) + \psi_\alpha(u) \delta \phi^\alpha(u)) = \delta W.$$

It is evident that the functional must contain ϵ as a factor, and if we make a transformation,

$$\int_S^{(3)} (\chi^\alpha(u) \delta\pi_\alpha(u) - \psi_\alpha(u) \delta\phi^\alpha(u)) du = -\delta \left\{ U + \int_S^{(3)} \pi_\alpha(u) \chi^\alpha(u) du \right\} \\ = -\delta K[\phi^\alpha(u), \pi_\alpha(u)], \text{ say.}$$

Therefore
$$\chi^\alpha(u) = -\frac{\delta K}{\delta\pi_\alpha(u)}, \quad \psi_\alpha(u) = \frac{\delta K}{\delta\phi^\alpha(u)}. \quad (57)$$

The most general infinitesimal transformation is accordingly defined by the equations

$$\phi^{*\alpha}(u) - \phi^\alpha(u) = +\epsilon \frac{\delta K}{\delta\pi_\alpha(u)}, \quad \pi_\alpha^*(u) - \pi_\alpha(u) = \epsilon \frac{\delta K}{\delta\phi^\alpha(u)}, \quad (58)$$

with K an arbitrary functional of the $\phi^\alpha(u)$ and $\pi_\alpha(u)$, and ϵ an arbitrary infinitesimal quantity independent of the $\phi^\alpha(u)$ and $\pi_\alpha(u)$.

Canonical equations of motion

The transformation from the functions on a surface S to those on a neighbouring surface S' may be interpreted as an infinitesimal canonical transformation. Let S be defined by $x^\mu = x^\mu(u)$, and each point (u) be displaced to

$$x^\mu(u) + \delta x^\mu(u), \quad \text{with} \quad \delta x^\mu(u) = \epsilon \left(\frac{\partial x^\mu}{\partial u^r} a^r(u) + \frac{\partial x^\mu}{\partial w} a^n(u) \right), \quad (59)$$

i.e. ϵa^n in the direction of the normal, and ϵa^r in the direction of the co-ordinate u^r . Then

$$K = \int_S^{(3)} U_\nu^\mu(u) N_\mu(u) \delta x^\nu(u) du = \int_S^{(3)} (a^n(u) \mathcal{H}(u) + a^r(u) h_r(u)) du. \quad (60)$$

In particular,

$$\left\{ \begin{array}{l} a^n(u) = 1 \\ a^r(u) = 0 \end{array} \right\} \quad \text{and} \quad \left\{ \begin{array}{l} a^n(u) = 0 \\ a^1(u) = 1 \\ a^r(u) = 0 \quad (r \neq 1) \end{array} \right\} \text{etc.,}$$

lead to the canonical equations of motion.

Invariance of the canonical equations

THEOREM IV. *Under a canonical transformation the canonical equations retain their form, and the new Hamiltonians are $H + (\partial/\partial w) W$, $h_r + (\partial/\partial u^r) W$, expressed as functionals of the new variables.*

Proof. I may be regarded as a functional of the initial and final surfaces S_1 and S_2 , and of the functions $\phi^\alpha(u, S_1)$, $\phi^\alpha(u, S_2)$. Suppose that S_1 and $\phi^\alpha(u, S_1)$ are kept fixed, and S_2 varied only 'bodily', that is, δw and δu^r are independent of u . Then (35) becomes

$$\delta I = \int_{S_2}^{(3)} \pi_\alpha(u) \delta\phi^\alpha(u) du + H\delta w + h_r\delta u^r, \quad (61)$$

and the equations

$$\frac{\partial \pi_\alpha(u)}{\partial w} = \frac{\delta H}{\delta\phi^\alpha(u)}, \quad \frac{\partial \pi_\alpha(u)}{\partial u^r} = \frac{\delta h_r}{\delta\phi^\alpha(u)}$$

follow from the requirement that δI is a perfect differential. The remaining canonical equations follow from

$$\delta \left(I - \int_{S_1}^{(3)} \phi^z(u) \pi_z(u) du \right) = - \int_{S_2}^{(3)} \phi^z(u) \delta \pi_z(u) du + H \delta w + h_r \delta u^r. \quad (62)$$

This shows that the canonical equations follow from the boundary formula.

Now make a surface-dependent canonical transformation to new variables $\phi^{*z}(u), \pi_z^*(u)$. Since S_2 is varied

$$\int_{S_1}^{(3)} (\pi_z(u) \delta \phi^z(u) - \pi_z^*(u) \delta \phi^{*z}(u)) = -\delta W + \frac{\delta W}{\delta w} \delta w + \frac{\delta W}{\delta u^r} \delta u^r,$$

$$\text{and so} \quad \delta(I + W) = \int_{S_1}^{(3)} (\pi_z^*(u) \delta \phi^{*z}(u)) du + \left(H + \frac{\delta W}{\delta w} \right) \delta w + \left(h_r + \frac{\delta W}{\delta u^r} \right) \delta u^r. \quad (63)$$

The left-hand side is again a perfect differential, and the canonical equations follow as before.

Generalization of theorem IV

If the new variables $q^\beta(y), p_\beta(y)$ are defined on some other domain than the u^r , it is more convenient to employ equations which involve *functional* differentiation. The new equations are

$$\left. \begin{aligned} \frac{\delta q^\beta(y)}{\delta w(u)} &= -\frac{\delta \mathcal{H}^*(u)}{\delta p_\beta(y)}, & \frac{\delta p_\beta(y)}{\delta w(u)} &= \frac{\delta \mathcal{H}^*(u)}{\delta q^\beta(y)}, \\ \frac{\delta q^\beta(y)}{\delta u^r(u)} &= -\frac{\delta \mathcal{H}_r^*(u)}{\delta p_\beta(y)}, & \frac{\delta p_\beta(y)}{\delta u^r(u)} &= \frac{\delta \mathcal{H}_r^*(u)}{\delta q^\beta(y)}, \end{aligned} \right\} \quad (64)$$

$$\text{where} \quad \mathcal{H}^*(u) = \mathcal{H}(u) + \frac{\delta W}{\delta w(u)}, \quad \mathcal{H}_r^*(u) = \mathcal{H}_r(u) + \frac{\delta W}{\delta u^r(u)}. \quad (65)$$

This theorem may be proved by generalizing the argument above, and will be needed later. Equations of the type (64) are appropriate because the new variables are functionals of the whole surface S .

Integrability conditions

The functional derivatives $(\delta/\delta w(u)) W$, $(\delta/\delta u^r(u)) W$ of any functional W of a surface S satisfy certain integrability conditions, analogous to those satisfied by the derivatives ϕ_i of a function ϕ of several variables, $(\partial/\partial x^k) \phi_i = (\partial/\partial x^i) \phi_k$, but rather more complicated, since the operators $(\partial/\partial x^k)$ all commute, while operators of functional differentiation in general do not (Dirac 1948, equations (35) to (37)). These conditions will be denoted by (A). Any four functions $(W_n(u), W_r(u))$ must certainly satisfy (A) if they are to be the functional derivatives of a functional W of S .

In quantum mechanics the state-vector $|\Psi, S\rangle$ is a functional of S , and according to Schrödinger's equations its functional derivatives are proportional to

$$\mathcal{H}(u) |\Psi, S\rangle, \quad \mathcal{H}_r(u) |\Psi, S\rangle.$$

By using (A), Dirac deduced from these expressions a necessary set of conditions in P.B. form, (B), which any set of Hamiltonian functions $(\mathcal{H}(u), \mathcal{H}_r(u))$ must satisfy in order to provide a consistent scheme of quantum mechanics (Dirac 1948, equations (49) to (51)).

One would expect to deduce the same conditions (B) classically by reasoning in a similar way from the Hamilton-Jacobi equations discussed below, since these are the analogues of Schrödinger's equations. Here the functional derivatives of W are actually equal to $(\mathcal{H}(u), \hbar_r(u))$, apart from their sign, which shows that (B) can only be a re-statement of (A) in P.B. form.

The classical Hamiltonian functions are, however, the functional derivatives of the action integral I , so that they certainly satisfy (A) and hence (B). Therefore any form of quantum mechanics which is deduced originally from a classical Lagrangian must be consistent. This provides a strong reason for believing that the Lagrangian is really more important in quantum mechanics than the Hamiltonian, although at present its significance is obscured.

The analogue in field dynamics of a time-dependent canonical transformation is one which depends on S , e.g. the transformation from the canonical variables on a fixed S_0 , to those on a variable S , obtained by solving the equations of motion. In transformations of this kind the functional W is a functional of S .

A surface-dependent canonical transformation might be defined not by W itself but by a set of four functions asserted to be the functional derivatives of W , together with some initial conditions. It would then be built up by integration. Such a set of functions must satisfy (A). The most important example is the transformation to the interaction representation, defined by the 'free' Hamiltonian functions $\mathcal{H}^f, \hbar_r^f$; here again (A) hold, because these functions are derived from the Lagrangian $\mathcal{L}^f$.

If a set of Hamiltonian functions satisfies (B), then after a canonical transformation has been performed the new Hamiltonian functions will again satisfy (B). For by the generalization to theorem IV, the new Hamiltonian functions differ from the old by the functional derivatives of W , which themselves satisfy (A) and therefore (B). One would of course expect a consistent theory of canonical transformations to possess this property. In particular, it shows that the interaction Hamiltonian function satisfies (B), and therefore that the interaction representation is consistent.

This argument will be carried through more thoroughly in the second paper. The present notation differs from that of Dirac (1948) in several details.

The Hamilton-Jacobi equations

The solution of the classical problem consists in determining $2n$ relations $F_a[\phi^\alpha(u), \pi_\alpha(u)] = \omega_a(v)$ ($a = 1, \dots, 2n$), where the $\omega_a(v)$ are a set of independent constant functions, and the F_a are functionals or, alternatively, $2n$ relations

$$\phi^\alpha(u) = f^\alpha[\omega_a(v)], \quad \pi_\alpha(u) = g_\alpha[\omega_a(v)].$$

These are called integrals of the system. The ten classical integrals, and the total electromagnetic charge, are not integrals in this sense, because they involve constants rather than constant functions. The problem is solved if one can find an S -dependent canonical transformation, generated by a functional $W[\phi^\alpha(u), q^\beta(v), S]$, from $\phi^\alpha(u), \pi_\alpha(u)$ to a new set $q^\beta(v), p_\beta(v)$ which does not depend on S . The equations

$$\pi_\alpha(u) = -(\delta/\delta\phi^\alpha(u)) W, \quad p_\beta(v) = (\delta/\delta q^\beta(v)) W \quad (66)$$

provide $2n$ integrals of the system.

The new variables are independent of S if the new Hamiltonian functions are identically zero, i.e. if

$$\mathcal{H}(u) + (\delta/\delta w(u)) W = 0, \quad \hbar_r(u) + (\delta/\delta u^r(u)) W = 0.$$

The problem therefore reduces to the following: determine any functional

$$W[\phi^\alpha(u), q^\beta(v), S]$$

involving n arbitrary constant functions $q^\beta(v)$ which satisfies the simultaneous partial functional differential equations

$$\left. \begin{aligned} \mathcal{H}\left(\phi^\alpha(u), \phi_r^\alpha(u), -\frac{\delta W}{\delta \phi^\alpha(u)}\right) + \frac{\delta W}{\delta w(u)} &= 0, \\ \hbar_r\left(\phi^\alpha(u), \phi_r^\alpha(u), -\frac{\delta W}{\delta \phi^\alpha(u)}\right) + \frac{\delta W}{\delta u^r(u)} &= 0. \end{aligned} \right\} \quad (67)$$

Then (66) provide $2n$ integrals of the system. (67) are called the Hamilton-Jacobi equations. Their chief interest here comes from the fact that they are very similar in form to Schrödinger's equations in quantum mechanics.

Perturbation theory

A perturbation theory may be constructed which is exactly similar to that in particle dynamics. Suppose that $(\mathcal{H}, \hbar_r)$ may be divided into two parts,

$$\mathcal{H} = \mathcal{H}^0 + \mathcal{H}^i, \quad \hbar_r = \hbar_r^0 + \hbar_r^i, \quad (68)$$

each set satisfying (B). The interesting case, used in setting up the interaction representation, is when $(\mathcal{H}^0, \hbar_r^0)$ are actually the Hamiltonian functions of a simpler problem whose Lagrangian is $\mathcal{L}^0$; then (B) are certainly satisfied, as we have seen. For the present we make this assumption. $(\mathcal{H}^i, \hbar_r^i)$ then represent an *interaction*, or perturbation of a system which, if left to itself, would have the simpler Hamiltonian functions $(\mathcal{H}^0, \hbar_r^0)$. In order to make the division (68), $\mathcal{H}$ and $\hbar_r$ must be expressed in terms of $\phi^\alpha(u)$, $\phi_r^\alpha(u)$ and $\pi_\alpha(u)$.

THEOREM V. $\hbar_r^i$ is necessarily identically zero.

Proof. The proof is immediate. For according to equations (23), $\hbar_r(u)$ has the same form for any problem. $\hbar_r$ and $\hbar_r^0$ are therefore identical and $\hbar_r^i$ zero.

Choose some initial surface S , and by a canonical transformation to be described construct a set of new variables $\phi^{*\alpha}(v, T; S)$, $\pi_\alpha^*(v, T; S)$, with the properties:

- (i) They are functions of the three parameters v^r of a variable space-like surface T .
- (ii) They depend on the initial surface S , but are not functions of the u^r .
- (iii) If T coincides with S , so that not only are the two surfaces τ, σ identical, but also the parameters v^r, u^r then

$$\left. \begin{aligned} \phi^{*\alpha}(v, T; S) &= \phi^\alpha(u, S), \\ \pi_\alpha^*(v, T; S) &= \pi_\alpha(u, S). \end{aligned} \right\} \quad (69)$$

(iv) They are obtained from the $\phi^\alpha(u)$, $\pi_\alpha(u)$ by a T -dependent canonical transformation generated by a functional $P[T, S]$ which satisfies

$$\frac{\delta P[T, S]}{\delta \rho(v)} = \mathcal{H}^0(*, v), \quad \frac{\delta P[T, S]}{\delta v^r(v)} = \hbar_r^0(*, v) \quad (70)$$

for variations of T , and therefore

$$\frac{\delta P[T, S]}{\delta w(u)} = -\mathcal{H}^0(*, u, S), \quad \frac{\delta P[T, S]}{\delta u^r(u)} = -h_r^0(*, u, S) \quad (71)$$

for variations of S , on account of (69). ρ is the normal co-ordinate of $\mathcal{T}$. $\mathcal{H}^0(*, v)$ means the function $\mathcal{H}^0$ of the new variables $\phi^{*\alpha}(v, T; S)$, $\pi_\alpha^*(v, T; S)$, evaluated at the point (v) on T . $\mathcal{H}^0(*, u, S)$ means the function $\mathcal{H}^0$ of the new variables, evaluated at the point $(v) = (u)$ on the surface $T = S$.

When S is fixed, and T varied, the new variables satisfy the canonical equations

$$\left. \begin{aligned} \frac{\delta \phi^{*\alpha}(v)}{\delta \rho} &= -\frac{\delta H^0[*]}{\delta \pi_\alpha^*(v)}, & \frac{\delta \pi_\alpha^*(v)}{\delta \rho} &= \frac{\delta H^0[*]}{\delta \phi^{*\alpha}(v)}, \\ \frac{\delta \phi^{*\alpha}(v)}{\delta v^r} &= -\frac{\delta h_r^0[*]}{\delta \pi_\alpha^*(v)}, & \frac{\delta \pi_\alpha^*(v)}{\delta v^r} &= \frac{\delta h_r[*]}{\delta \phi^{*\alpha}(v)}. \end{aligned} \right\} \quad (72)$$

For the old variables did not depend on T , but only on S , so that the corresponding Hamiltonians were zero. By theorem IV, under a T -dependent canonical transformation generated by P , the new Hamiltonians are formed from the old by adding the quantities $\frac{\delta P}{\delta \rho}$, $\frac{\delta P}{\delta v^r}$.

If T is fixed and S varied the new variables satisfy the canonical equations

$$\left. \begin{aligned} \frac{\delta \phi^{*\alpha}(v, T; S)}{\delta w(u)} &= -\frac{\delta \mathcal{H}^i(*, u, S)}{\delta \pi_\alpha^*(v, T; S)}, & \frac{\delta \phi^{*\alpha}(v, T; S)}{\delta u^r(u)} &= 0, \\ \frac{\delta \pi_\alpha^*(v, T; S)}{\delta w(u)} &= \frac{\delta \mathcal{H}^i(*, u, S)}{\delta \phi^{*\alpha}(v, T; S)}, & \frac{\delta \pi_\alpha^*(v, T; S)}{\delta u^r(u)} &= 0, \end{aligned} \right\} \quad (73)$$

which follow from the generalization of theorem IV, and from (71). This type of canonical equation is correct because the new variables are functionals of the whole surface S , and not just functions defined at a point on it.

By (73) the new variables depend only on the σ associated with S , and not on the parametrization of S . We may use them to define new space-time variables $\phi^{*\alpha}(x; \sigma)$, satisfying the Lagrangian equations

$$\frac{\partial \mathcal{L}^0(*)}{\partial \phi^{*\alpha}} - \frac{\delta}{\delta x^\mu} \left(\frac{\partial \mathcal{L}^0(*)}{\partial \phi_\mu^{*\alpha}} \right) = 0, \quad (74)$$

and also $\phi^{*\alpha}(x; \sigma) = \phi^\alpha(x)$, $n_\mu (\partial / \partial \phi_\mu^{*\alpha}) \mathcal{L}^0(*) = n_\mu (\partial / \partial \phi_\mu^\alpha) \mathcal{L}^0$

for all points (x) on σ , since n_μ is proportional to N_μ . The parametrization does not appear in this prescription. In the new problem the Hamilton-Jacobi equations for variations of S are

$$\frac{\delta W[S]}{\delta w(u)} + \mathcal{H}^i(*, u, S) = 0, \quad \frac{\delta W[S]}{\delta u^r(u)} = 0, \quad (75)$$

so that $W[S]$ does not depend on the parametrization of S , and (75) could be written $(\delta / \delta \sigma(x)) W[\sigma] + \mathcal{H}^i(*, x, \sigma) = 0$ ((x) a point on σ).

By solving (74) one may construct invariant P.B. relations between the $\phi^{*z}(x; \sigma)$ at two points of space-time. The P.B. equations of motion can be written

$$\frac{\delta F[* , \tau]}{\delta x^\nu} = [F[* , \tau], P_\nu^0[* , \tau]], \quad (76)$$

where $F[* , \tau]$ is a dynamical variable which, considered as a function or functional of the canonical variables, does not explicitly depend on T , and the energy-momentum vector is defined by

$$P_\nu^0[* , \tau] = \int_T^{(3)} U_\nu^{0\mu}(*) N_\mu dv. \quad (77)$$

The $\phi^{*z}(x; \sigma)$ and their space and time derivatives are included in this category. P_ν^0 does not depend on the parametrization, which together with the canonical functions has completely disappeared from the theory.

The interaction representation may be constructed by choosing for $\mathcal{L}^0$ the free Lagrangian $\mathcal{L}^f$, and taking this perturbation scheme over into quantum mechanics, using the Heisenberg picture for variations of (x) and the Schrödinger picture for those of σ . The dynamical variables then satisfy equations of motion corresponding to (74) or (76), and do not depend on σ , while the state-vector satisfies an equation similar to (75) and does not depend on (x) . This procedure will be discussed in detail in part II.

$\mathcal{L}^f$ is not the only possible choice for $\mathcal{L}^0$; if an external electromagnetic field were present for instance, one might include the extra part of the Lagrangian in $\mathcal{L}^0$, instead of treating the field as a perturbation. This method would be a good one if the modified equations (74) could be solved exactly, as they could for a Coulomb field; the effect of the field would then be to modify the invariant P.B. relations. The radiative perturbations of atomic energy-levels could be calculated in this way. The general rule is that $\mathcal{L}^0$ should be the Lagrangian of a simple problem which can be solved. (Throughout the paper it has been tacitly assumed that the Lagrangian does not depend explicitly on (x) , to simplify the notation, but there is no need to do so, and external fields can be included without difficulty.)

A perturbation treatment can be employed even if $(\mathcal{H}^0, \hbar_r^0)$ in (68) are not the Hamiltonian functions of a simpler problem; for instance, if this transformation has already been carried out. But it is then necessary to ensure that (B) are satisfied by constructing the generating functional P explicitly. In this way a problem may be solved by successive canonical transformations, each generated by some suitable functional P . This method is already included in the general discussion of canonical transformations; it is the classical analogue of the scheme used by Schwinger (1949*a*), and has no special theory attached to it of the kind described above.

REFERENCES

- Belinfante, F. J. 1949 *Phys. Rev.* **76**, 66.
 Chang, T. S. 1949 *Phys. Rev.* **75**, 4967.
 Dirac, P. A. M. 1948 *Phys. Rev.* **73**, 1092.
 Kanesawa, S. & Tomonaga, S. 1948*a* *Progr. Theor. Phys.* **3**, 1.
 Kanesawa, S. & Tomonaga, S. 1948*b* *Progr. Theor. Phys.* **3**, 101.

- Koba, Z., Tati, T. & Tomonaga, S. 1947*a* *Progr. Theor. Phys.* **2**, 101.
Koba, Z., Tati, T. & Tomonaga, S. 1947*b* *Progr. Theor. Phys.* **2**, 198.
Matthews, P. T. 1949 *Phys. Rev.* **76**, 1270.
Schwinger, J. 1948 *Phys. Rev.* **74**, 1439.
Schwinger, J. 1949*a* *Phys. Rev.* **75**, 651.
Schwinger, J. 1949*b* *Phys. Rev.* **76**, 790.
Tomonaga, S. 1946 *Progr. Theor. Phys.* **1**, 27.
Weiss, P. 1936 *Proc. Roy. Soc. A*, **156**, 192.
Weiss, P. 1938 *Proc. Roy. Soc. A*, **169**, 102.

Wave equations in momentum space

BY MAURICE LÉVY, DR.-ÈS-SC.,* *University of Manchester*†

(Communicated by R. E. Peierls, F.R.S.—Received 13 May 1950)

The non-relativistic and relativistic wave equations of a particle are solved in momentum space for different kinds of fields. It is assumed that the potential, when defined in co-ordinate space, is central; but the method applies also to more general forms of momentum space potentials.

First, the general integral equations which correspond, in momentum space, to Schrödinger and Dirac equations are derived, and the angular variables separated.

Then, a general method of solution of these integral equations is given, which lies mainly on a transformation into a four-dimensional hyperspace. These results are (relativistically and non-relativistically) applied rigorously to the Coulombian field (hydrogen atom) and to the scalar Yukawa potential.

A more general case is further investigated. In this, the momentum space potential can be expanded in a series of powers of the modulus of the momentum (this last application involves a cut-off procedure).

A mathematical appendix deals with some properties of the hyperspherical harmonics (especially the integral equation of these functions) and with the Fourier transform of some kinds of central potentials.

I. INTRODUCTION

The momentum representation has rarely been used in the study of atomic or nuclear problems. This representation is generally believed to lead to great mathematical difficulties, probably because integral, instead of differential, equations are involved in the formulation of problems. However, there are many questions in atomic and nuclear physics which are much simplified by this representation. Typical examples are the hydrogen atom (see §IV) and the binding energies of light nuclei (Svartholm 1945). In addition, the physical significance of some quantities is more directly apparent in momentum space (especially in collision problems). On the other hand, the momentum distribution in atoms, which is important in several connexions, can always be obtained more simply by a direct solution of the wave equation in momentum space than by a Fourier transformation of the solution in co-ordinate space.

The momentum representation has recently found an important application in the study of the relativistic interaction between nucleons (Van Hove 1949). This approach to the meson theory of nuclear forces eliminates the 'spurious' singularity at the origin of the static meson potential. The interaction Hamiltonian is an integral operator in both co-ordinate and momentum spaces, but it is much simpler in the momentum representation. The author is at present attempting to treat the deuteron problem using this relativistic Hamiltonian, and one of the principal difficulties is the absence of general methods of solving the wave equation in momentum space. It has therefore seemed worth while to treat the question from a general point of view and to give detailed examples of the application of the mathe-

* Chargé de recherches au Centre National de la Recherche Scientifique (Paris).

† Present address: The Institute for Advanced Study, Princeton (New Jersey), U.S.A.

mathematical technique. A general method of solving the wave equation in momentum space has been developed by generalizing Fock's method of transforming to a four-dimensional hyperspace (Fock 1935) which allows the expression of the solutions as series of hyperspherical harmonics.

The general formulation of the Schrödinger equation in momentum space has been made by Weyl (1928), while Gordon (1934) has investigated certain mathematical aspects of the question. Elsasser (1933) has applied this representation to the solution of certain elementary problems. But it is the hydrogen atom which has been most thoroughly studied.

In general, two distinct methods are available. The first, employed by Podolsky & Pauling (1929), depends on an application of the Fourier transformation to the solutions of the equation in co-ordinate space. The second attempts to formulate the problem directly in momentum space (as initiated by Dirac 1927). This, again, can be done in two different ways, both of which have been used in the case of the hydrogen atom: Hylleraas (1932) obtains, by an iteration method, the differential equation which defines the radial part of the solution. Fock (1935) starts from the integral equation which is derived by applying the Fourier transformation to the Cartesian Schrödinger equation, and shows that the 'accidental degeneracy' of the non-relativistic hydrogen atom is related to the invariance of the four-dimensional momentum space with respect to the rotation group. These two methods have yielded, in this special case, identical results. This is merely an accident, due to the degeneracy of the problem. In fact, the Fourier transformation is equivalent, in Hilbert functional space, to a finite rotation applied to the eigen-vectors in which the axis system does not remain invariant. Generally, linear combinations of the new eigen-vectors will be obtained. These do not necessarily correspond to an orientation of the axis system which is convenient when the degeneracy has to be removed.

While non-relativistic problems have been studied exhaustively, relativistic problems have been almost ignored. Only a short paper by Rubinowicz (1948), which will be discussed in §IV, can be quoted. But, if the momentum representation is to be used in relativistic problems, it is essential to be able to solve the Dirac equation in representation. Consequently, I have here tried to examine this problem as completely as possible, and, in this respect, it has seemed of interest to consider also the Schrödinger equation in each special case, in the light of the method which is used in this paper. In the following pages, only the case of a particle of mass μ in a static potential $V(r)$ will be investigated, but the method can be extended immediately to all the types of potentials which depend only on the modulus of the momentum. The calculations are made in the following way: in §II, the general wave equations in momentum representation are derived. A general method of solution of the momentum space integral equations is presented in §III. Then, the preceding results are applied to the problems of the non-relativistic and relativistic hydrogen atom (§IV). After that, in order to show the interest of the method and make possible a new discussion of the relativistic interactions between particles, the case of the scalar 'meson' potential is studied (§V), and the more general type of problems in which the momentum space potential can be expanded in a series of powers of a given parameter (§VI).

In this section we obtain successively the Schrödinger and Dirac equations in momentum space and separate the angular variables.

An elementary, but very instructive, study of the one-dimensional Schrödinger equation is given by Mott & Sneddon (1948, p. 68). The general problem in a $(q+2)$ -dimensional space is considered here.*

As stated in § 1, we only consider, in co-ordinate space, potential functions which depend on r alone. In this case, U depends on $|\mathbf{p} - \mathbf{p}'|$ alone, and equation (6) can be written as follows:

$$(p^2 - 2\mu E) \Phi(p, \theta_1, \theta_2, \dots, \theta_q, \phi) \\ = \frac{2\mu}{h^{\frac{1}{2}q+1}} \int_{(\omega')_{q+1}} U[(p^2 + p'^2 - 2pp' \cos \Theta)^{\frac{1}{2}}] \Phi(p', \theta'_1, \theta'_2, \dots, \theta'_q, \phi') d\omega'_{q+1}, \quad (8)$$

where we have put

$$d\omega'_{q+1} = \sin^q \theta'_1 \sin^{q-1} \theta'_2 \dots \sin \theta'_q d\theta'_1 d\theta'_2 \dots d\theta'_q d\phi, \quad (9)$$

Θ being the angle between $\mathbf{p}$ and $\mathbf{p}'$.

As the potential is central, the solutions are invariant under rotation around the origin. We can then separate the angular variables by putting

$$\Phi(\mathbf{p}) = G(\mathbf{p}) Y_l^{(q+1)}(\theta_1, \theta_2, \dots, \theta_q, \phi), \quad (10)$$

$Y_l^{(q+1)}$ being a spherical harmonic of degree l in the $(q+2)$ -dimensional space (see Appendix, subsection 1). We shall make use now of the theorem of Hecke (1918)—and its generalized form by Erdelyi (1938)—which defines the general type of integral equation of which the hyperspherical harmonics are solutions. This theorem will play a very important role throughout this paper. The reader is referred to the appendix (subsection 1) for mathematical details. Substituting (10) in (8), we obtain

$$(p^2 - 2\mu E) G(p) = \frac{2\mu}{h^{\frac{1}{2}q+1}} \int_0^{+\infty} G(p') K_l(p, p') p'^{q+1} dp', \quad (11)$$

with
$$K_l(p, p') = \frac{2\pi^{\frac{1}{2}(q+1)}}{\Gamma\left(\frac{q+1}{2}\right)} \int_{-1}^{+1} U[(p^2 + p'^2 - 2pp'x)^{\frac{1}{2}}] P_{l,0}^{(q)}(x) (1-x^2)^{\frac{1}{2}(q-1)} dx, \quad (12)$$

$P_{l,0}^{(q)}(x)$ being defined by (A 3).*

It is sometimes convenient to put $p_0 = \sqrt{(-2\mu E)}$, $p/p_0 = \rho$, and to write equation (11) in the form

$$(1 + \rho^2) G(\rho) = \frac{2\mu p_0^q}{h^{\frac{1}{2}q+1}} \int_0^{+\infty} G(\rho') K_l(\rho, \rho') \rho'^{q+1} d\rho'. \quad (13)$$

The argument of U in equation (12) is then replaced by $p_0(\rho^2 + \rho'^2 - 2\rho\rho'x)^{\frac{1}{2}}$.

The Schrödinger equation is thus reduced to a one-variable integral equation. Consequently, the problem can always be solved, and one can obtain approximate solutions of this equation by the Gauss-Hilbert variational method or the variation-iteration method of Svartholm (1945). Examples will be given in §§ IV and V of convergent solutions of this equation. However, it must be remembered that in most cases it will not be possible to construct convergent solutions, if the momenta vary in the range $(-\infty, +\infty)$. It is the case, for example, when $V(r)$ is of the form r^ν (cf. appendix, subsection 3). The reader will find an illustration of this case in § VI, where a more general type of solution is investigated.

* The equations of the appendix are numbered separately, their numbers being preceded by the letter A.

(2) Dirac equations

Only the four-dimensional case is investigated here. The Dirac equation can also be studied in a five-dimensional space (Dirac 1935), but it seems more useful to discuss the equation which is most generally used.

(a) Derivation of the integral equation

With the usual notation, we write the equation of a particle of mass μ in the form

$$\left(\boldsymbol{\alpha} \cdot \mathbf{p} + \beta \frac{E_0}{c} + p_4\right) \Psi(\mathbf{r}, \tau) = V(r) \Psi(\mathbf{r}, \tau). \quad (14)$$

$\mathbf{p}$ is a three-dimensional vector, whose components are p_1, p_2, p_3 ; similarly, $\boldsymbol{\alpha}$ has the components $\alpha_1, \alpha_2, \alpha_3$. We have put $E_0 = \mu c^2$. The component p_4 does not play a symmetrical role with respect to the others ($V(r)$ depends only on the Cartesian *spatial* co-ordinates), and the method of derivation of the momentum space equation is somewhat different from the method which we have used in the case of the Schrödinger equation. The Fourier transformation changes $\Psi(\mathbf{r}, \tau)$ into $\Pi(\mathbf{p}, p_4)$, τ being a co-ordinate proportional to the time.* The equation (14) can then be written

$$\left(\boldsymbol{\alpha} \cdot \mathbf{p} + \beta \frac{E_0}{c} + p_4\right) \Pi(\mathbf{p}, p_4) = \frac{1}{h^2} \iiint_{-\infty}^{+\infty} V(r) \Psi(x, y, z, \tau) e^{i(\mathbf{p} \cdot \mathbf{r} + p_4 \tau)/\hbar} d\mathbf{r} d\tau. \quad (15)$$

The transformation of the potential $V(r)$ takes the form

$$W(\mathbf{p}, p_4) = \frac{1}{h^2} \iiint_{-\infty}^{+\infty} V(r) e^{i(\mathbf{p} \cdot \mathbf{r} + p_4 \tau)/\hbar} d\mathbf{r} d\tau, \quad (16)$$

and consequently

$$W(\mathbf{p}, p_4) = \frac{1}{h} \delta(p_4) \int_{(\mathbf{r})} V(r) e^{i(\mathbf{p} \cdot \mathbf{r})/\hbar} d\mathbf{r} = h^{\frac{1}{2}} \delta(p_4) U(\mathbf{p}), \quad (17)$$

where $U(\mathbf{p})$ is the three-dimensional Fourier transform of the co-ordinate potential. Accordingly, we can write the Dirac equation in the following way:

$$\left(\boldsymbol{\alpha} \cdot \mathbf{p} + \beta \frac{E_0}{c} + p_4\right) \Pi(\mathbf{p}, p_4) = \frac{1}{h^{\frac{1}{2}}} \int_{(p'_4)} \int_{(\mathbf{p}')} U(|\mathbf{p} - \mathbf{p}'|) \delta(p_4 - p'_4) \Pi(\mathbf{p}', p'_4) d\mathbf{p}' dp'_4. \quad (18)$$

One can get rid of the δ function which appears on the right-hand side by putting

$$\Pi(\mathbf{p}, p_4) = \Phi(\mathbf{p}) H(p_4), \quad (19)$$

and $p_4 = E/c$. One obtains then the following three-dimensional equation:

$$\left[\boldsymbol{\alpha} \cdot \mathbf{p} + \frac{1}{c} (E + \beta E_0)\right] \Phi(\mathbf{p}) = \frac{1}{h^{\frac{1}{2}}} \int_{(\mathbf{p}')} U(\mathbf{p} - \mathbf{p}') \Phi(\mathbf{p}') d\mathbf{p}', \quad (20)$$

* One can take, for example, $\tau = ict$.

(b) *Separation of the angular variables*

In order to separate the angular variables, we write the four components of Φ as expressions proportional to spherical harmonics of the type $Y_l^\mu(\theta, \phi)$.^{*} By again applying the theorem of Hecke (equation (A 1, 2)), we obtain, on the right-hand side of (20) and for each component, an expression proportional to Y_l^μ . The solutions have then, in momentum space, the same form as in co-ordinate space.[†] If, furthermore, we make use of the multiplication rules by $\sin \theta e^{\pm i\phi}$ and $\cos \theta$ (Bethe 1933, p. 551), we obviously get, as far as angular dependence is concerned, expressions analogous in all points to those which are used in the theory of the hydrogen atom in co-ordinate space. In particular, there are two types of solutions, corresponding respectively to lower indices l and $l+1$ ($j = l + \frac{1}{2}$) or l and $l-1$ ($j = l - \frac{1}{2}$). We only consider here the case where $j = l + \frac{1}{2}$; the other one can be deduced from this one without any difficulty, l being replaced by $l-1$. We are then induced to put

$$\left. \begin{aligned} \Phi_1(\mathbf{p}) &= \left(\frac{l+m+1}{2l+3} \right)^{\frac{1}{2}} Y_{l+1}^m(\theta, \phi) M(\rho), \\ \Phi_2(\mathbf{p}) &= - \left(\frac{l-m+2}{2l+3} \right)^{\frac{1}{2}} Y_{l+1}^{m-1}(\theta, \phi) M(\rho), \\ \Phi_3(\mathbf{p}) &= \left(\frac{l-m+1}{2l+1} \right)^{\frac{1}{2}} Y_l^m(\theta, \phi) N(\rho), \\ \Phi_4(\mathbf{p}) &= \left(\frac{l+m}{2l+1} \right)^{\frac{1}{2}} Y_l^{m-1}(\theta, \phi) N(\rho). \end{aligned} \right\} \quad (21)$$

with $\rho = p/p_0$, $p_0 = (1/c) E_0(1-\epsilon^2)^{\frac{1}{2}}$, and $\epsilon = E/E_0$.

We deduce, for M and N , the following system of integral equations:

$$\left. \begin{aligned} \Gamma M(\rho) + \rho N(\rho) &= \frac{1}{h^{\frac{3}{2}}} \int_0^{+\infty} M(\rho') K_{l+1}(\rho, \rho') \rho'^2 d\rho', \\ \Gamma \rho M(\rho) - N(\rho) &= \frac{\Gamma}{h^{\frac{3}{2}}} \int_0^{+\infty} N(\rho') K_l(\rho, \rho') \rho'^2 d\rho', \end{aligned} \right\} \quad (22)$$

where we have put, by application of Hecke's theorem,

$$\left. \begin{aligned} K_l(\rho, \rho') &= 2\pi p_0^2 \int_{-1}^{+1} U[p_0(\rho^2 + \rho'^2 - 2\rho\rho'x)^{\frac{1}{2}}] P_l(x) dx \\ \text{and} \quad \Gamma &= [(1+\epsilon)/(1-\epsilon)]^{\frac{1}{2}}. \end{aligned} \right\} \quad (23)$$

If we consider the definition of ϵ , we see that Γ is large compared to unity. Consequently, the system (22) shows the separation between the large component $N(\rho)$ and the small one $M(\rho)$. We have thus reduced the general problem to the resolution of a system of linear integral equations, solutions of which can always be found by approximate methods.

^{*} The normalization conditions are defined in the appendix. In the case of three dimensions they agree with those of Bethe (1933).

[†] The physical meaning of this result (which, mathematically, follows from Hecke's theorem) lies in the fact that the Fourier transformation leaves invariant the total angular momentum and the product $(\mathbf{m} \cdot \boldsymbol{\sigma})$ of the orbital momentum and the spin angular momentum (cf. Rubinowicz 1948).

(c) *Reduction of the system*

Sometimes it is convenient to express the equations (22) as a system of two Schrödinger equations relating separately to the functions M and N . Let us extract $M(\rho)$ from the second equation and introduce it in the first one. We obtain

$$(1 + \rho^2) N(\rho) + \frac{\Gamma}{h^{\frac{3}{2}}} \int_0^{+\infty} N(\rho') K_l(\rho, \rho') \rho'^2 d\rho' = \frac{1}{\Gamma h^{\frac{3}{2}}} \int_0^{+\infty} \frac{\rho}{\rho'} N(\rho') K_{l+1}(\rho, \rho') \rho'^2 d\rho' + \frac{1}{h^3} \int_0^{+\infty} \int_0^{+\infty} \frac{\rho}{\rho''} N(\rho') K_{l+1}(\rho, \rho'') K_l(\rho'', \rho') \rho''^2 \rho'^2 d\rho' d\rho''. \quad (24)$$

Next we do the same for $N(\rho)$ and we put

$$\left. \begin{aligned} \Delta_l^{(1)}(\rho, \rho') &= K_{l+1}(\rho, \rho') + \frac{\Gamma}{h^{\frac{3}{2}}} \int_0^{+\infty} \frac{\rho}{\rho''} K_l(\rho, \rho'') K_{l+1}(\rho'', \rho') \rho''^2 d\rho'' - \Gamma^2 \frac{\rho}{\rho'} K_l(\rho, \rho'), \\ \Delta_l^{(2)}(\rho, \rho') &= -K_l(\rho, \rho') + \frac{1}{\Gamma h^{\frac{3}{2}}} \int_0^{+\infty} \frac{\rho}{\rho''} K_{l+1}(\rho, \rho'') K_l(\rho'', \rho') \rho''^2 d\rho'' + \frac{1}{\Gamma^2} \frac{\rho}{\rho'} K_{l+1}(\rho, \rho'). \end{aligned} \right\} \quad (25)$$

After some calculations we obtain the system

$$\left. \begin{aligned} (1 + \rho^2) M(\rho) &= \frac{1}{\Gamma h^{\frac{3}{2}}} \int_0^{+\infty} M(\rho') \Delta_l^{(1)}(\rho, \rho') \rho'^2 d\rho', \\ (1 + \rho^2) N(\rho) &= \frac{\Gamma}{h^{\frac{3}{2}}} \int_0^{+\infty} N(\rho') \Delta_l^{(2)}(\rho, \rho') \rho'^2 d\rho'. \end{aligned} \right\} \quad (26)$$

We must keep in mind, however, that these equations are the result of an iteration of the system (22), which is submitted to restrictive conditions. For example, the limits of integration in equation (24) are infinite, and it is not always allowed to reverse the order of integration. The greatest care must be taken in this operation to ensure that the solutions of the iterated system (26) are solutions of the initial system (22).

(d) *Method of resolution by successive approximations*

When the kernels $\Delta_l^{(1)}$ and $\Delta_l^{(2)}$ of the preceding system are too difficult to calculate, and when the exact or approximate solution of the corresponding Schrödinger equation is known, a useful method consists in calculating successive approximations starting from the separation between large and small components.

In the first approximation, let us put $M(\rho') = 0$ on the right-hand side of the first equation of system (22). We obtain the system

$$\left. \begin{aligned} M_1(\rho) &= \frac{\rho}{\Gamma} N_1(\rho), \\ (1 + \rho^2) N_1(\rho) &= \frac{\Gamma}{h^{\frac{3}{2}}} \int_0^{+\infty} N_1(\rho') K_l(\rho, \rho') \rho'^2 d\rho'. \end{aligned} \right\} \quad (27)$$

This last equation is the Schrödinger equation corresponding to the same potential. If its solutions are known, we can, in principle, calculate without any difficulty the successive approximations of the functions M and N . We obtain a series of functions which are the coefficients of successive powers of $1/\Gamma$ which is generally a very small quantity.

III. GENERAL METHOD OF RESOLUTION OF THE RADIAL EQUATIONS BY TRANSFORMATION IN A FOUR-DIMENSIONAL MOMENTUM SPACE

All the integral equations with which we have to deal in this paper contain integrals of the type which appears on the right-hand side of (13) and (22). The solution of these equations has been found to be more apparent if the variables are transformed according to a general transformation initiated by Fock (1935).

Let us, for example, consider equation (13) (with $q = 1$), which defines the radial part of the solutions of the Schrödinger equation in a three-dimensional space. With the combination of definition (12) it may be written

$$(1 + \rho^2) G(\rho) = \frac{2\mu p_0}{h^{\frac{3}{2}}} 2\pi \int_{\rho'=0}^{+\infty} \int_{x=-1}^{+1} G(\rho') U[p_0(\rho^2 + \rho'^2 - 2\rho\rho'x)^{\frac{1}{2}}] P_{l,0}^{(1)}(x) \rho'^2 d\rho' dx. \quad (28)$$

The transformation of Fock is obtained by putting

$$\rho = \tan \frac{\psi}{2}, \quad \rho' = \tan \frac{\psi'}{2}. \quad (29)$$

This is equivalent to transforming the momentum space ($\mathcal{E}_m$) into a sphere Σ of radius p_0 imbedded in a four-dimensional hyperspace ($\mathcal{E}_f$). ($\mathcal{E}_m$) is deduced from the hypersphere of space ($\mathcal{E}_f$) simply by a stereographical projection.

Moreover, we put $x = \cos \theta'$. The integral of the right-hand side of (28) can easily be transformed into a surface integral on Σ , with three variables ψ', θ', ϕ' :

$$\frac{G(\psi)}{\cos^2 \frac{1}{2}\psi} = \frac{\mu p_0}{4h^{\frac{3}{2}}} \int_0^\pi \int_0^\pi \int_0^{2\pi} \frac{G(\psi')}{\cos^6 \frac{1}{2}\psi'} U\left[\frac{p_0 \sin \frac{1}{2}\Theta}{\cos \frac{1}{2}\psi \cos \frac{1}{2}\psi'}\right] P_{l,0}^{(1)}(\cos \theta') d\Omega', \quad (30)$$

with

$$d\Omega' = \sin^2 \psi' \sin \theta' d\psi' d\theta' d\phi'. \quad (31)$$

Θ is the angle between two unit vectors $\mathbf{u}$ and $\mathbf{v}$ which, in the space ($\mathcal{E}_f$), are defined respectively by the polar angles $(\psi, 0, 0)$ and (ψ', θ', ϕ') .

The solutions of equation (30) will be expressed, in general, by a combination of hyperspherical harmonics of the space ($\mathcal{E}_f$) by means of the theorem of Hecke and a suitable use of the recurrence formulae (A 7) to (A 9). The simplest example will be given below for the non-relativistic hydrogen atom where $U(z) = z^{-2}$. In this case no use of the recurrence formulae is needed, and the solution is expressed only in terms of *one* hyperspherical harmonic. In order to illustrate the more general aspects of the method, we shall consider here the case in which $U(z) = z^{-3}$. Here, equation (30) transforms into

$$\frac{G(\psi)}{\cos^5 \frac{1}{2}\psi} = \frac{\mu}{4p_0^2 h^{\frac{3}{2}}} \int_0^\pi \int_0^\pi \int_0^{2\pi} \frac{G(\psi')}{\cos^3 \frac{1}{2}\psi'} \frac{1}{\sin^3 \frac{1}{2}\Theta} P_{l,0}^{(1)}(\cos \theta') d\Omega'. \quad (32)$$

$$\text{We make the substitution} \quad G(\psi) = \cos^5 \frac{1}{2}\psi H(\psi), \quad (33)$$

$$\text{and obtain} \quad H(\psi) = \frac{\mu}{8h^{\frac{3}{2}} p_0^2} \int_0^\pi \int_0^\pi \int_0^{2\pi} \frac{(1 + \cos \psi') H(\psi')}{\sin^3 \frac{1}{2}\Theta} P_{l,0}^{(1)}(\cos \theta') d\Omega'. \quad (34)$$

Let us imagine that $H(\psi')$ contains a term of the form $P_{n,l}^{(2)}(\cos \psi')$. The multiplication by $(1 + \cos \psi')$ on the right-hand side will lead to the appearance of the three harmonics $P_{n \pm 1, l}^{(2)}, P_{n, l}^{(2)}$. Then, if we attempt a solution of the form

$$H(\psi) = \sum_k a_k P_{n+k}^{(2)}(\cos \psi), \quad (35)$$

equation (34), with the application of the theorem of Hecke, transforms into a set of linear equations:

$$a_k - \frac{\mu}{8\hbar^2 p_0^2} \left[a_k + \frac{(n+k+1)(n+k-l)}{2(n+k)^2} a_{k-1} + \frac{(n+k+1)(n+k+l+2)}{2(n+k+2)^2} a_{k+1} \right] \Lambda_{n+k} = 0, \quad (36)$$

for all values of k . In these equations Λ_{n+k} is defined by

$$\frac{1}{\Lambda_{n+k}} = \pi \sqrt{2} \int_{-1}^{+1} \frac{1}{(1-x)^{\frac{3}{2}}} P_{n+k}^{(2)}(x) (1-x^2)^{\frac{1}{2}} dx. \quad (37)$$

Equations (46) allow the calculation of all successive terms of the expansion (35). More general cases, in which the kernel $U(z)$ is itself expanded in a series of powers of a given parameter, will be presented in §§ V and VI.

IV. THE HYDROGEN ATOM

We shall consider, as a consequence of our general study of § II, the case in which $V(r)$ is the 'Coulombian' potential of a hydrogen-like atom

$$V = \frac{e^2 Z}{r}, \quad (38)$$

e being the electronic charge, Z the atomic number of the atom considered. In subsection 3 of the appendix, we have calculated the Fourier transform of the potential V in a $(q+2)$ -dimensional space. In the case in which $q = 1$, we get (equation A 18)

$$U(\mathbf{p}) = 2\pi\hbar^{\frac{1}{2}} e^2 Z \frac{1}{|\mathbf{p}|^2}. \quad (39)$$

We shall investigate in turn the non-relativistic and the relativistic problems.

(1) *Non-relativistic problem (Schrödinger equation)*

With the special type of potential with which we are concerned, we may write equation (13), for the radial part $G(\rho)$ of the solutions, in the following form:

$$(1+\rho^2) G(\rho) = \frac{\mu e^2 Z}{2\pi^3 \hbar p_0} \int_0^{+\infty} K_l(\rho, \rho') G(\rho') \rho'^2 d\rho', \quad (40)$$

with
$$K_l(\rho, \rho') = 2\pi \int_{-1}^{+1} \frac{P_l(x) dx}{\rho^2 + \rho'^2 - 2\rho\rho'x}. \quad (41)$$

The kernel $K_l(\rho, \rho')$ can be calculated with the help of Neumann's formula (Whittaker & Watson 1927, p. 320):

$$K_l(\rho, \rho') = \frac{2\pi}{\rho\rho'} Q_l\left(\frac{\rho^2 + \rho'^2}{2\rho\rho'}\right). \quad (42)$$

In fact, it is more convenient, in order to solve equation (40), to retain the integral representation of $K_l(\rho, \rho')$ and to transform the integral of the second term of (41) by means of the general method of the last section. Consequently, we go over into the hyperspace ($\mathcal{C}_f$) by means of the substitution $\rho = \tan \frac{1}{2}\psi$. Let us put

$$G(\psi) = \cos^4 \frac{1}{2}\psi H(\psi). \quad (43)$$

Equation (40) may be written

$$H(\psi) = \frac{\lambda}{2\pi} \int_0^\pi \int_0^\pi \int_0^{2\pi} \frac{H(\psi') P_l(\cos \theta') d\Omega'}{4 \sin^2 \frac{1}{2} \Theta}, \quad (44)$$

with $\lambda = (\mu e^2 Z)/\pi \hbar p_0$. Application of Hecke's theorem gives the solution of equation (44)

$$H(\psi) = P_{n,l}^{(2)}(\cos \psi), \quad (45)$$

the energy levels being determined by the relation $\lambda = 2\pi/\Lambda_n$, with

$$\Lambda_n = \frac{(2\pi)^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} \int_{-1}^{+1} \frac{1}{2(1-x)} P_{n,0}^{(2)}(x) (1-x^2)^{\frac{1}{2}} dx. \quad (46)$$

Taking into account the definition of $P_{n,0}^{(p)}(x)$ (cf. appendix, subsection 1) and the relation (Magnus & Oberhettinger 1948, p. 99),

$$C_n^1(\cos \theta) = \frac{\sin(n+1)\theta}{\sin \theta}, \quad (47)$$

we obtain

$$\Lambda_n = \frac{2\pi}{n+1} \int_0^{+\infty} \frac{1}{1-\cos \theta} \sin(n+1)\theta \sin \theta d\theta = \frac{2\pi^2}{n+1}. \quad (48)$$

The energy levels equation can thus be written*

$$\frac{\mu e^2 Z}{\hbar p_0} = n+1. \quad (49)$$

Using the general definition of hyperspherical harmonics (Magnus & Oberhettinger 1948) and equations (12), (43) and (45), we obtain the following general solution of the equations:

$$\begin{aligned} \Phi_{n,l,m}(\psi, \theta, \phi) \\ = \frac{1}{\sqrt{2}} \frac{\Gamma(n+1) \Gamma(l+\frac{3}{2}) \Gamma(n+l+2)}{\Gamma(n+2) \Gamma(2l+2) \Gamma(n-l+1)} \sin^{-\frac{1}{2}} \psi \cos^{\frac{1}{2}} \psi P_{n+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \psi) P_l^m(\cos \theta) e^{im\phi}. \end{aligned} \quad (50)$$

If, using the analytical expression (42) of the nucleus, one expresses that $P_{n,l}^{(2)}(\cos \psi)$ is a solution of equation (44), one obtains with the old variables ρ and ρ' the interesting formula

$$\begin{aligned} \rho^{\frac{1}{2}}(1+\rho^2)^{-\frac{1}{2}} P_{n+\frac{1}{2}}^{-(l+\frac{1}{2})} \left(\frac{1-\rho^2}{1+\rho^2} \right) \\ = \frac{\sqrt{2}(n+1)}{\pi} \int_0^{+\infty} \rho'(1+\rho'^2)^{-2} P_{n+\frac{1}{2}}^{-(l+\frac{1}{2})} \left(\frac{1-\rho'^2}{1+\rho'^2} \right) Q_l \left(\frac{\rho^2+\rho'^2}{2\rho\rho'} \right) d\rho'. \end{aligned} \quad (51)$$

We must note that Fock's space is not the only hyperspace in which the solutions of equation (40) may be expressed by means of hyperspherical harmonics. Let us put, for example,

$$\rho = \tan \eta. \quad (52)$$

We transform, in this way, the momentum space ($\mathcal{E}_m$) into a sphere of radius p_0 embedded in a hyperspace ($\mathcal{E}_n$) of which ($\mathcal{E}_m$) is the orthogonal projection. The passage of ($\mathcal{E}_f$) to ($\mathcal{E}_n$) is done by the transformation

$$\psi = 2\eta. \quad (53)$$

* The number n used by Fock is, with our definition of hyperspherical harmonics, equal to $n+1$.

We now use the formula (A 6) which allows us to express the Legendre functions of $\cos 2\theta$ by means of those the argument of which is $\cos \theta$. We get, from equation (50),

$$\Phi_{n,l,m}(\psi, \theta, \phi) = \frac{1}{\sqrt{2}} \frac{\Gamma(n+1) \Gamma(l+\frac{3}{2}) \Gamma(n+l+2)}{\Gamma(n+2) \Gamma(2l+2) \Gamma(n-l+1)} \sum_{k=0}^{n-l-1} \mathcal{F}_{k,l,m}(\eta, \theta, \phi), \quad (54)$$

$$\text{with } \mathcal{F}_{k,l,m}(\eta, \theta, \phi) = \frac{(-2)^k \Gamma(n-l)}{k! \Gamma(n-l-k)} \cos^k \eta P_{k+l+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \eta) P_l^m(\cos \theta) e^{im\phi}. \quad (55)$$

Equation (55) expresses the fact that the transformation (53) is equivalent, in the functional Hilbert space where the functions $\Phi_{n,l,m}$ are represented, to a finite rotation applied to the eigen vectors. These vectors are expressed, after the transformation, as a linear combination of the initial eigen vectors. The passage from the $\Phi_{n,l,m}$ to $\mathcal{F}_{n,l,m}$ will be useful in the interpretation of some results of the next subsection.

(2) Relativistic problem (Dirac equation)

We have given in §II (2) a general form of the solutions in the case of a central potential. We need only, in the particular case with which we are concerned here, to calculate the two radial functions of the solutions. These functions satisfy (with the notation of §II) the system of equations

$$\left. \begin{aligned} \rho N(\rho) + \Gamma M(\rho) &= \frac{\alpha Z}{\pi} \int_0^{+\infty} M(\rho') K_{l+1}(\rho, \rho') d\rho', \\ -N(\rho) + \Gamma \rho M(\rho) &= \frac{\alpha Z \Gamma}{\pi} \int_0^{+\infty} N(\rho') K_l(\rho, \rho') d\rho'. \end{aligned} \right\} \quad (56)$$

The kernel $K_l(\rho, \rho')$ is defined by (41) and (42); α is the fine structure constant: $e^2/\hbar c = 1/137$.

(a) Solution in hyperspace ($\mathcal{E}_f$)

In order to solve the system (56), we go over into the hyperspace ($\mathcal{E}_f$) as usual, by means of Fock's transformation (29). Moreover, we put

$$\left. \begin{aligned} M(\psi) &= \cos^4 \frac{1}{2} \psi R(\psi), \\ N(\psi) &= \cos^4 \frac{1}{2} \psi S(\psi); \end{aligned} \right\} \quad (57)$$

the system (56) takes the form

$$\left. \begin{aligned} \Gamma(1 + \cos \psi) R(\psi) + \sin \psi S(\psi) &= \frac{\alpha Z}{2\pi} \int_0^\pi \int_0^\pi \int_0^{2\pi} \frac{R(\psi') P_{l+1}(\cos \theta') d\Omega'}{4 \sin^2 \frac{1}{2} \Theta}, \\ \Gamma \sin \psi R(\psi) - (1 + \cos \psi) S(\psi) &= \frac{\alpha Z \Gamma}{2\pi} \int_0^\pi \int_0^\pi \int_0^{2\pi} \frac{S(\psi') P_l(\cos \theta') d\Omega'}{4 \sin^2 \frac{1}{2} \Theta}. \end{aligned} \right\} \quad (58)$$

The expression on the right-hand side of equations (58) is very similar to the corresponding expression of the Schrödinger equation in hyperspace ($\mathcal{E}_f$) (equation (54)). We are then induced to attempt a solution of the following type:

$$\left. \begin{aligned} R(\psi) &= \sum_{k=0}^{+\infty} a_k \sin^{-\frac{1}{2}} \psi P_{\nu+k+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \psi), \\ S(\psi) &= \sum_{k=0}^{+\infty} b_k \sin^{-\frac{1}{2}} \psi P_{\nu+k+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \psi). \end{aligned} \right\} \quad (59)$$

Using the theorem of Hecke as in the first part of this section, and recurrence relations (A 19), (A 20) and (A 21), the system (58) transforms into

$$\left. \begin{aligned} \sum_{k=-1}^{+\infty} \left\{ \frac{\nu+k+l+2}{2(\nu+k)} [\Gamma a_{k-1} + b_{k-1}(\nu+k+l+1)] + a_k \left(\Gamma - \frac{\alpha Z}{\nu+k+1} \right) \right. \\ \left. + \frac{\nu+k+l}{2(\nu+k+2)} [\Gamma a_{k+1} - b_{k+1}(\nu+k-l+1)] \right\} P_{\nu+k+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \psi) = 0, \\ \sum_{k=-1}^{+\infty} \left\{ \frac{1}{2(\nu+k)} [\Gamma a_{k-1} + b_{k-1}(\nu+k+l+1)] + b_k \left(1 + \frac{\alpha Z \Gamma}{\nu+k+1} \right) \right. \\ \left. - \frac{1}{2(\nu+k+2)} [\Gamma a_{k+1} - b_{k+1}(\nu+k-l+1)] \right\} P_{\nu+k+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \psi) = 0. \end{aligned} \right\} \quad (60)$$

Equating separately to zero the coefficients of $P_{\nu+k+\frac{1}{2}}^{-(l+\frac{1}{2})}$ and $P_{\nu+k+\frac{1}{2}}^{-(l+\frac{1}{2})}$, we get a series of linear equations, which allows the calculation of all the coefficients successively. In order to find quickly the relation of compatibility of these equations (which gives the energy levels), we put first $k = -1$ in equations (60), and we remark that $a_{-2} = b_{-2} = a_{-1} = b_{-1} = 0$. The system of equations (60) thus reduces to

$$\Gamma a_0 - b_0(\nu - l) = 0. \quad (61)$$

Then we put $k = 0$:

$$\left. \begin{aligned} a_0 \left(\Gamma + \frac{\alpha Z}{\nu+1} \right) + \frac{\nu-l}{2(\nu+2)} [\Gamma a_1 - b_1(\nu-l+1)] = 0, \\ -b_0 \left(1 + \frac{\alpha Z \Gamma}{\nu+1} \right) + \frac{1}{2(\nu+2)} [\Gamma a_1 - b_1(\nu-l+1)] = 0. \end{aligned} \right\} \quad (62)$$

In the last two equations, a_1 and b_1 appear only through the sum $c_0 = \Gamma a_1 - b_1(\nu-l+1)$. Equations (61) and (62) can consequently be considered as a linear homogeneous system in a_0, b_0, c_0 . Its compatibility relation is

$$\alpha Z \Gamma^2 - 2(\nu+1) \Gamma - \alpha Z = 0. \quad (63)$$

If we continue this process, putting $k = 1$ in equations (60), we obtain two new equations, in which a_2 and b_2 appear through the expression $c_1 = \Gamma a_2 - b_2(\nu-l+2)$; hence, we have a system of five linear and homogeneous equations in a_0, b_0, a_1, b_1, c_1 . It can be shown that the corresponding compatibility determinant, and all the determinants which can be obtained in continuing this process, are equal to zero if condition (63) is satisfied. The positive root of this equation can be written

$$\Gamma = \frac{\nu+1 + \sqrt{[(\nu+1)^2 + \alpha^2 Z^2]}}{\alpha Z}, \quad (64)$$

$$\text{and consequently} \quad \epsilon = \frac{E}{m_0 c^2} = \frac{\nu+1}{\sqrt{[(\nu+1)^2 + \alpha^2 Z^2]}}. \quad (65)$$

We obtain the usual expression for the energy levels if we take

$$\nu = n' + \gamma - 1. \quad (66)$$

The solutions of system (58) have then the form

$$\left. \begin{aligned} M(\psi) &= \cos^4 \frac{1}{2} \psi \sin^{-\frac{1}{2}} \psi \sum_{k=1}^{+\infty} a_k P_{n'+\gamma+k+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \psi), \\ N(\psi) &= \cos^4 \frac{1}{2} \psi \sin^{-\frac{1}{2}} \psi \sum_{k=1}^{+\infty} b_k P_{n'+\gamma+k+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \psi). \end{aligned} \right\} \quad (67)$$

This expression of the solutions is convenient, because the large component converges very quickly. The order of magnitude of Γ is $2(\nu+1)/\alpha Z$, then very large. Equation (62) gives

$$\lambda_0 = \frac{b_0(\nu-l)}{\Gamma} \approx \frac{b_0(\nu-l)\alpha Z}{\alpha(\nu+1)}, \quad (68)$$

and expresses the separation into large and small components, all the ratios of the coefficients a_k and b_k being of the order of αZ . In the same way, the successive ratios b_{k+1}/b_k are of the order of $(\alpha Z)^2$. For example, one has, for the first term

$$b_1 \approx \frac{\alpha^2 Z^2 (\nu+2)(3\nu+5)b_0}{2(\nu+1)^2(\nu-l+1)}. \quad (69)$$

(b) Solutions in hyperspace ($\mathcal{E}_n$)

Rubinowicz (1948) has obtained the functions M and N by direct application of the Fourier transformation to the solutions of co-ordinate space. In the case in which $j = l + \frac{1}{2}$, he obtains the following expressions:

$$\left. \begin{aligned} M(\rho) &= B \sqrt{(1-\epsilon)} \{ n' G(n'-1, 2\gamma+1, \gamma-1, l+1, \rho) \\ &\quad + (N+l+1) G(n', 2\gamma+1, \gamma-1, l+1, \rho) \}, \\ N(\rho) &= -B \sqrt{(1+\epsilon)} \{ -n' G(n'-1, 2\gamma+1, \gamma-1, l, \rho) \\ &\quad + (N+l+1) G(n', 2\gamma+1, \gamma-1, l, \rho) \}, \end{aligned} \right\} \quad (70)$$

the function G being defined by

$$G(\nu, \beta, \delta, l, \rho) = \frac{1}{\rho^{\frac{1}{2}}} \int_0^{+\infty} e^{-t} t^{\delta+\frac{1}{2}} {}_1F_1(-\nu, \beta, t) J_{l+\frac{1}{2}}\left(\frac{\rho t}{2}\right) dt. \quad (71)$$

With these notations, which are those of Bethe (1933), n' and N are respectively the radial and total quantum numbers; γ is equal to $[(l+1)^2 - \alpha^2 Z^2]^{\frac{1}{2}}$. For the definition of B , the reader is referred to the original paper. Let us consider, for the sake of simplicity, the case in which $|\rho| \leq 1$. In this case, Rubinowicz gets the following representation:

$$\begin{aligned} G(\nu, \beta, \delta, l, \rho) &= \frac{\nu! \Gamma(\beta)}{\Gamma(l+\frac{1}{2})} 2^{\delta-l+2} \rho \sum_{n=0}^{\nu} \frac{\Gamma(n+l+\delta+3)}{n! \Gamma(\nu-n+1) \Gamma(\beta+n)} \\ &\quad \times (-2)^n F\left(\frac{n+l+\delta+3}{2}, \frac{n+l+\delta+4}{2}, l+\frac{3}{2}, -\rho^2\right). \end{aligned} \quad (72)$$

We shall transform the hypergeometric function which appears under the summation index, in order to express it by means of the hyperspherical harmonics of space ($\mathcal{E}_n$)

which have been introduced in the first part of this section. By means of the transformation (52), and using the definition of general Legendre functions of the second kind, given by Hobson (1931, p. 196), we get

$$G(\nu, \beta, \delta, l, \tan \eta) = \frac{\nu! \Gamma(\beta)}{\sqrt{\pi}} 2^{\delta+3} e^{-i(\delta+l+\frac{1}{2})\pi} \cot^l \eta \cos^{\delta+2} \eta \\ \times \sum_{n=0}^{\nu} \frac{2^n \Gamma(n+l+\delta+3)}{n! \Gamma(\nu-n+1) \Gamma(\beta+n) \Gamma(n-l+\delta+1)} \cos^n \eta Q_l^{-(n+\delta+1)}(i \cot \eta). \quad (73)$$

We now use the relation between general Legendre functions of the first and second kind, given by Hobson (1931, p. 247). Expression (73) may then be written

$$G = \frac{\nu! \Gamma(\beta)}{\sqrt{2}} 2^{\delta+3} e^{-i(2\delta+l+1)\pi} \sum_{n=0}^{\nu} G_n, \quad (74)$$

with

$$G_n = \frac{(-2)^n \Gamma(n+l+\delta+3)}{n! \Gamma(\nu-n+1) \Gamma(\beta+n)} \cos(l+\delta+1) \frac{1}{2} \eta \sin^{-\frac{1}{2}} \eta \cos^n \eta P_{n+\delta+\frac{1}{2}}^{-(l+\frac{1}{2})}(\cos \eta). \quad (75)$$

It can be proved, starting from the expression which Rubinowicz gives for $|\rho| > 1$, that equation (74) is valid for all values of ρ . If we compare it to the solution of the Schrödinger equation in space ($\mathcal{E}_n$), given by equation (54), we see that, apart from the normalizing coefficient, the two expressions differ only by a very small variation of the indices. In the case where $j = l + \frac{1}{2}$, $\delta = \gamma - 1$, δ differs from l only by a quantity of the order of $\alpha^2 Z^2$.

The expression of solutions in space ($\mathcal{E}_n$) involves only a finite number of hyper-spherical harmonics, instead of an infinite number in space ($\mathcal{E}_f$). But, for practical purposes, the latter seems to be more convenient as the large component converges more quickly.*

V. THE 'MESON' POTENTIAL

We turn now to the case of a particle of mass μ , placed in a central scalar 'meson' potential

$$V(r) = g^2 \frac{e^{-kr}}{r}. \quad (76)$$

The Fourier transformation of $V(r)$ is equal to (cf. appendix subsection 3)

$$U(\mathbf{p}) = \frac{4\pi g^2 \hbar^{\frac{1}{2}}}{|\mathbf{p}|^2 + k^2 \hbar^2}. \quad (77)$$

We shall examine mainly the Schrödinger equation, which, as will be seen, can be used, with this special type of potential, for various kinds of problems. Of the Dirac equation, whose field of application is more limited, we shall say only a few words.

* As this paper was being sent to the Editor, my attention was drawn to a paper by Rubinowicz (*Prace Mat.-Fiz.* 1949, 47, 1-8) in which the author obtains the solutions of the relativistic one-electron problem by transforming the integral equation into a set of differential equations. This procedure is, in its principle, different from the method of Hylleraas mentioned in § 1. It can be described as a transformation leading back to ordinary space, confined to the radial part of the solution.

(1) *Schrödinger equation*

According to the general theory which we have given in §II we need only to consider the radial part $G(\rho)$ of the solutions ($\rho = p/p_0$, with $p_0 = (-2\mu E)^{1/2}$). We obtain the equation

$$(1 + \rho^2) G(\rho) = \frac{\lambda}{2\pi} \int_0^{+\infty} G(\rho') K_l(\rho, \rho') \rho'^2 d\rho', \quad (78)$$

with $\lambda = 16\pi^2 \mu g^2 / \hbar p_0$, and

$$K_l(\rho, \rho') = 2\pi \int_0^\pi \frac{P_l(\cos \theta') \sin \theta' d\theta'}{\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta' + \chi^2}, \quad (79)$$

where we have put $\chi = k\hbar/p_0$.

In order to solve the integral equation, we shall use a method of resolution by successive approximations similar to the perturbation method, expanding the kernel $K_l(\rho, \rho')$ in a series of powers of χ . However, the greatest care must be taken in performing this expansion, for a too superficial procedure would lead to a series of diverging integrals.

Let us consider, in a three-dimensional space with $Oxyz$ as system of reference, a vector $\mathbf{p} = \mathbf{OM}$ the length of which is equal to ρ , and which is situated along Oz , and another vector $\mathbf{p}' = \mathbf{OM}'$, whose longitude and colatitude are ϕ' and θ' . In these conditions we have

$$(\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta')^{1/2} = |\mathbf{p} - \mathbf{p}'|. \quad (80)$$

If we take into account the fact that the expansion of $(1+x)^{-1}$ in powers of x is convergent* only if $|x| \leq 1$, we see that $K_l(\rho, \rho')$ must be expanded in the following way

$$K_l(\rho, \rho') = 2\pi \sum_{p=0}^{+\infty} (-1)^p \chi^{2p} K_l^{(p)}(\rho, \rho'), \quad (81)$$

with

$$K_l^{(p)}(\rho, \rho') = \int_0^\pi \frac{P_l(\cos \theta') \sin \theta' d\theta'}{(\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta')^{p+1}}, \quad (82)$$

$$\text{if } |\mathbf{p} - \mathbf{p}'| \geq \chi, \text{ and } K_l(\rho, \rho') = 2\pi \sum_q (-1)^q \frac{1}{\chi^{2q+2}} K_l^{(q)}(\rho, \rho'), \quad (83)$$

$$\text{with } K_l^{(q)}(\rho, \rho') = \int_0^\pi (\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta')^q P_l(\cos \theta') \sin \theta' d\theta', \quad (84)$$

if $|\mathbf{p} - \mathbf{p}'| \leq \chi$.

We can, as in §§ III and IV, replace the integral of the second member of (78) by a volume integral extended to the whole three-dimensional space Ω' . We isolate out of this space a small sphere Ω_0 whose centre is M and radius χ . Then, the equation (78) may be written

$$(1 + \rho^2) G(\rho) = \frac{\lambda}{2\pi} \sum_p (-1)^p \chi^{2p} \int_{(\Omega' - \Omega_0)} \frac{G(\rho') P_l(\cos \theta') d\mathbf{p}'}{(\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta')^{p+1}} \\ + \frac{\lambda}{2\pi} \sum_q (-1)^q \frac{1}{\chi^{2q+2}} \int_{\Omega_0} (\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta')^q G(\rho') P_l(\cos \theta') d\mathbf{p}'. \quad (85)$$

In order to calculate the second term of the right-hand side of (85), we assume that χ is a small quantity. When the particle under consideration is a nucleon, and the

* When $x = 1$, the series is only semi-convergent (its sum depends on the order of the terms), but this fact is of little importance here.

Yukawa potential results from an interaction by π mesons of mass μ_0 , this condition can be written

$$\chi = \frac{\mu_0}{(-2\mu E)^{\frac{1}{2}}} \ll 1, \quad (86)$$

and, consequently, the binding energy must be such that $|E| \gg \mu_0^2/2\mu \approx 9.75 \text{ MeV}$. More generally, for a particle of mass μ_0 and energy E , the range of the potential must fulfil the condition* $\mu_0 \gg \hbar(-2\mu E)^{-\frac{1}{2}}$. With these assumptions, it seems natural to replace, in the second integral of the right-hand side of (85), $G(\rho') P_l(\cos \theta')$ by its value at the centre of the sphere Ω_0 , that is to say, $G(\rho)$. If we adopt a reduced system of polar co-ordinates r_0, θ_0, ϕ_0 , with M as a pole, this term can be written

$$\frac{\lambda}{2\pi} \sum_{q=0}^{+\infty} (-1)^q \frac{1}{\chi^{2q+1}} G(\rho) \int_0^\chi \int_0^\pi \int_0^{2\pi} r_0^{2q+2} \sin \theta_0 d\theta_0 d\phi_0 = 2\lambda \chi G(\rho) \sum_{q=0}^{+\infty} (-1)^q \frac{1}{2q+3}. \quad (87)$$

The value of the series which appears on the right-hand side of (87) is $1 - \frac{1}{4}\pi$. The equation (85) can then be written

$$(1 + \rho^2 - \epsilon\chi) G(\rho) = \frac{\lambda}{2\pi} \sum_{p=0}^{+\infty} (-1)^p \chi^{2p} \int_{(\Omega' - \Omega_0)} \frac{G(\rho') P_l(\cos \theta') d\mathbf{p}'}{(\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta')^{p+1}}, \quad (88)$$

where we have put $\epsilon = 2\lambda(1 - \frac{1}{4}\pi)$.

We shall now solve the equation (88) by successive approximations. Consequently, we expand $G(\rho)$ in series of the powers of χ ,

$$G = G_0 + \chi G_1 + \chi^2 G_2 + \dots \quad (89)$$

(a) Zero-order approximation

In the approximation, we neglect all the terms the order of which is superior or equal to χ . The equation (88) can accordingly be written

$$(1 + \rho^2) G_0(\rho) = \frac{\lambda}{2\pi} \int_{(\Omega' - \Omega_0)} \frac{G_0(\rho') P_l(\cos \theta') d\mathbf{p}'}{\rho^2 + \rho'^2 - 2\rho\rho' \cos \theta'}. \quad (90)$$

If we go over into the hyperspace ($\mathcal{E}_f$), we see that, as for the Coulombian case, the solution is

$$G_0(\psi) = \cos^4 \frac{1}{2}\psi P_{n,0}^{(2)}(\cos \psi). \quad (91)$$

The levels are determined by a slightly different condition $\lambda = 4\pi/\Lambda_n^0$, with

$$\Lambda_n^0 = \frac{2\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} \int_{-1}^{1-\epsilon_0} \frac{1}{1-x} P_{n,0}^{(2)}(x) (1-x^2)^{\frac{1}{2}} dx, \quad (92)$$

ϵ_0 being defined by† $\epsilon_0 = 2\chi^2 \cos^4 \frac{1}{2}\psi$.

* With this condition it seems difficult to use the method that follows for nuclear problems. But this procedure can be applied without much difficulty to atomic or molecular problems: One can mention, for example, the capture of a light particle by a complex atom and the 'screening effect' of the atomic electrons shells. As for nuclear problems, it remains to be seen if equation (78) cannot be solved by another expansion procedure which would not require condition (86).

† This condition arises from the fact that $|\mathbf{p} - \mathbf{p}'|$ is equal, in the hyperspace ($\mathcal{E}_f$) to $(1 - \cos \Theta)/2 \cos^2 \frac{1}{2}\psi \cos^2 \frac{1}{2}\psi'$ (cf. §III, p. 13) and the condition (84) gives, for the limiting value,

$$1 - \cos \Theta_0 = 2\chi \cos^2 \frac{1}{2}\psi \cos^2 \frac{1}{2}\psi'$$

from which follows (92) if one notices that at this point, $\cos^2 \frac{1}{2}\psi'$ differs from $\cos^2 \frac{1}{2}\psi$ only by a term in χ^2 .

We give here an expansion of Λ_n^0 up to the term in χ^2 , which will also be useful in the next approximation:

$$\Lambda_n^0(\chi, \psi) = \frac{2\pi^2}{n+1} - 8\pi\chi \cos^2 \frac{1}{2}\psi + \text{term in } \chi^3 \dots \quad (93)$$

(b) *First-order approximation*

Up to the terms in χ^2 , we may write equation (88) as follows:

$$(1 + \rho^2 - \epsilon\chi) G_0 + \chi(1 + \rho^2) G_1 = \frac{\lambda}{2\pi} \int_{(\Omega' - \Omega_0)} \frac{(G_0 + \chi G_1) P_l(\cos \theta') d\mathbf{p}'}{|\mathbf{p} - \mathbf{p}'|^2} - \frac{\lambda}{2\pi} \chi^2 \int_{(\Omega' - \Omega_0)} \frac{G_0(\rho') P_l(\cos \theta') d\mathbf{p}'}{|\mathbf{p} - \mathbf{p}'|^4}. \quad (94)$$

Taking into account (91) and (93), this can be written in space ($\mathcal{E}_f$)

$$\begin{aligned} \chi R_1(\psi) - \frac{\chi\lambda}{4\pi} \int_{(S' - S_0)} \frac{R_1(\psi') P_l(\cos \theta') d\Omega'}{1 - \cos \Theta'} \\ = \left[\frac{\lambda}{4\pi} \frac{1}{\Lambda_n^0} + \epsilon\chi \cos^2 \frac{1}{2}\psi - 1 \right] P_{n,l}^{(2)}(\cos \psi) - F_1(\psi), \end{aligned} \quad (95)$$

where we have put

$$F_1(\psi) = \frac{\lambda}{4\pi} \chi^2 \cos^2 \frac{1}{2}\psi \int_{(S' - S_0)} \frac{P_{n,l}^{(2)}(\cos \psi') P_l(\cos \theta') \cos^2 \frac{1}{2}\psi' d\Omega'}{(1 - \cos \Theta')^2}, \quad (96)$$

$$R_1(\psi) = \frac{G_1(\psi)}{\cos^4 \frac{1}{2}\psi}, \quad (97)$$

and where S' and S_0 are the portions of the surface of the hypersphere in which Ω' and Ω_0 are transformed.*

In order to calculate $F_1(\psi)$, we first evaluate the expression

$$(\cos^2 \frac{1}{2}\psi) P_{n,l}^{(2)}(\cos \psi) = \frac{1}{2}(1 + \cos \psi) P_{n,l}^{(2)}(\cos \psi)$$

which, taking into account the definition of the hyperspherical harmonics, we may write

$$\begin{aligned} \cos^2 \frac{1}{2}\psi P_{n,l}^{(2)}(\cos \psi) \\ = \frac{1}{2} \left[P_{n,l}^{(2)}(\cos \psi) + \frac{(n+2)(n-l+1)}{2(n+1)^2} P_{n+1,l}^{(2)}(\cos \psi) + \frac{n(n+l+1)}{2(n+1)^2} P_{n-1,l}^{(2)}(\cos \psi) \right]. \end{aligned} \quad (98)$$

Consequently, we obtain

$$\begin{aligned} F_1(\psi) = \frac{\lambda}{8\pi} \chi^2 \cos^2 \frac{1}{2}\psi \left[\Lambda_n^{(1)}(\chi, \psi) P_{n,l}^{(2)}(\cos \psi) \right. \\ + \Lambda_{n+1}^{(1)}(\chi, \psi) \frac{(n+2)(n-l+1)}{2(n+1)^2} P_{n+1,l}^{(2)}(\cos \psi) \\ \left. + \Lambda_{n-1}^{(1)}(\chi, \psi) \frac{n(n+l+1)}{2(n+1)^2} P_{n-1,l}^{(2)}(\cos \psi) \right], \end{aligned} \quad (99)$$

* $F_1(\psi)$ seems to be of χ^2 order; but, as we shall see, the integral on the right-hand side of (96) behaves like χ^{-1} , and $F_1(\psi)$ is actually of the first order in χ .

where we have put

$$\Lambda_n^{(1)}(\chi, \psi) = \frac{(2\pi)^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} \int_{-1}^{1-\epsilon_0} \frac{1}{(1-x)^2} P_{n,0}^{(2)}(x) (1-x^2)^{\frac{1}{2}} dx. \quad (100)$$

In this approximation, we only need the term in χ^{-1} in the expansion of $\Lambda_n^{(1)}(\chi, \psi)$:

$$\Lambda_n^{(1)}(\chi, \psi) \approx \frac{16\pi}{\alpha_0} = \frac{8\pi}{\chi \cos^2 \frac{1}{2}\psi}. \quad (101)$$

If we bring this value of $\Lambda_n^{(1)}$ into (99), and if we make use of relation (98) in order to calculate the $(\cos^2 \frac{1}{2}\psi) P_{n,l}^{(2)}(\cos \psi)$ term which appears in the right-hand side, we see that this side of the equation becomes a linear combination of the three harmonics $P_{n,l}^{(2)}(\cos \psi)$ and $P_{n\pm 1,l}^{(2)}(\cos \psi)$. We are then induced to write*

$$R_1(\psi) = a_1 P_{n+1,l}^{(2)}(\cos \psi) + a_{-1} P_{n-1,l}^{(2)}(\cos \psi). \quad (102)$$

The integral of the first member of (95) has then the value

$$\frac{\lambda\chi}{4\pi} [a_1 \Lambda_{n+1}^{(0)} P_{n+1,l}^{(2)}(\cos \psi) + a_{-1} \Lambda_{n-1}^{(0)} P_{n-1,l}^{(2)}(\cos \psi)]. \quad (103)$$

We obtain the energy levels by equating the coefficient of $P_{n,l}^{(2)}(\cos \psi)$ to zero. Taking into account (98) and (101), we obtain

$$\frac{\lambda\pi}{2(n+1)} \approx 1 + \frac{2\chi}{\pi} (n+1) \left(1 + \frac{\pi}{4}\right) + \dots \quad (104)$$

The coefficients a_1 and a_{-1} have the value (limited to the first order terms)

$$\left. \begin{aligned} a_1 &= -\frac{1}{\pi} \left(1 + \frac{1}{4}\pi\right) \frac{(n+2)^2 (n-l+1)}{n+1}, \\ a_{-1} &= \frac{1}{\pi} \left(1 + \frac{1}{4}\pi\right) \frac{n^2 (n+l+1)}{n+1}. \end{aligned} \right\} \quad (105)$$

To the first order in χ , the function G is equal, in space $(\mathcal{C}_f)$, to

$$G^{(1)}(\psi) = \cos^4 \frac{1}{2}\psi \left\{ P_{n,l}^{(2)}(\cos \psi) + \frac{\chi}{\pi(n+1)} \left(1 + \frac{1}{4}\pi\right) [n^2 (n+l+1) P_{n-1,l}^{(2)}(\cos \psi) - (n+2)^2 (n-l+1) P_{n+1,l}^{(2)}(\cos \psi)] \right\}. \quad (106)$$

We must note that the system of functions $G^{(1)}$ is not completely orthogonal. Only the functions $G^{(1)}(n)$ and $G^{(1)}(n')$ such that $|n - n'| > 2$ are orthogonal. But it is always possible to orthogonalize such a system (and, more generally, any system of approximate values of functions G) by means of the method of Schmidt.†

In a general way it can be seen that the coefficient of χ^r in the expansion of $G(\psi)/\cos^4 \frac{1}{2}\psi$ is a linear combination of the two hyperspherical harmonics $P_{n-r,l}^{(2)}(\cos \psi)$ and $P_{n+r,l}^{(2)}(\cos \psi)$. We can write then

$$G(\psi) = \cos^4 \frac{1}{2}\psi \left\{ P_{n,l}^{(2)}(\cos \psi) + \sum_{r=1}^{+\infty} \chi^r [a_r(n, l) P_{n-r,l}^{(2)}(\cos \psi) + b_r(n, l) P_{n+r,l}^{(2)}(\cos \psi)] \right\}. \quad (107)$$

* The introduction of a term proportional to $P_{n,l}^{(2)}(\cos \psi)$ in $R_1(\psi)$ would lead to a term in χ^2 in equation (94), which is impossible at this stage of the calculation.

† See, for example, Lichnerowicz (1947, p. 204).

(2) Dirac equation

From the general investigation of § II it can be seen that the system of Dirac radial equations has, in the present case, to be written in the following form:

$$\left. \begin{aligned} \rho N(\rho) - \Gamma M(\rho) &= \frac{\lambda}{2\pi} \int_0^{+\infty} M(\rho') K_{l+1}(\rho, \rho') \rho'^2 d\rho', \\ N(\rho) + \Gamma \rho M(\rho) &= \frac{\lambda \Gamma}{2\pi} \int_0^{+\infty} N(\rho') K_l(\rho, \rho') \rho'^2 d\rho', \end{aligned} \right\} \quad (108)$$

where we have put $\lambda/2\pi = g^2/\hbar^2 p_0$, and where $K_l(\rho, \rho')$ is defined by (79). An analytical form of this kernel can be given by means of Neumann's formula (Whittaker & Watson 1927, p. 330):

$$K_l(\rho, \rho') = \frac{2\pi}{\rho\rho'} Q_l\left(\frac{\rho^2 + \rho'^2 + \chi^2}{2\rho\rho'}\right), \quad (109)$$

Q_l being the Legendre function of the second kind. In order to solve the system of equations (116), it is possible to generalize the method of the last section (expansion of the nucleus in powers of χ), starting from the solutions of the relativistic Coulombian problem, given in § IV. Approximate solutions can also be obtained by means of the method outlined at the end of § II, which is based on the separation between large and small components. However, we shall not enter into further details, as the scalar static potential (76) is not correct for relativistic nuclear problems, and therefore is of very small interest.*

IV. DISCUSSION OF A MORE GENERAL TYPE OF POTENTIAL

In order to exhibit further some of the possibilities of the methods used throughout this paper, we investigate now the more general case, where the function $U(z)$, which defines the potential in momentum space, can be expanded in a series of powers of z , by means of a parameter ξ , sufficiently small to ensure a rapid convergence of the series. For the sake of simplicity, we only consider the Schrödinger equation in a three-dimensional space.

(1) Let us consider first the simplest case, where the series contains only even powers of z :

$$U(z) = a_0 + a_2 \xi^2 z^2 + a_4 \xi^4 z^4 + \dots \quad (110)$$

The kernel $K_l(p, p')$ can also be expanded in a series of functions of p and p' :

$$K_l(p, p') = \sum_r a_{2r} \xi^{2r} K_l^{(r)}(p, p'), \quad (111)$$

with
$$K_l^{(r)}(p, p') = 2\pi \int_{-1}^{+1} (p^2 + p'^2 - 2pp'x)^r P_l(x) dx. \quad (112)$$

As we have
$$\int_{-1}^{+1} x^r P_l(x) dx = 0 \quad \text{for } r < l, \quad (113)$$

* In an investigation which is in progress, we are extending the general method of § II to the relativistic deuteron problem, using, instead of potential (76), the relativistic interaction potentials obtained by Van Hove (1949).

$K_l^{(r)}$ differs from zero only for $r \geq l$. The first term of the nucleus (corresponding to approximation (l) for the function G) is then

$$K_l^{(l)}(p, p') = 2\pi \int_{-1}^{+1} (p^2 + p'^2 - 2pp'x)^l P_l(x) dx = 2\pi\gamma_l (-2pp')^l, \quad (114)$$

where we have put $\gamma_l = 2^{l+1}(l!)^2/(2l+1)!$. In the same way, we expand G in the form

$$G(p) = G_l(p) + \xi^2 G_{l+1}(p) + \dots \quad (115)$$

The Schrödinger equation for $G_l(p)$ can be written

$$(p^2 + p_0^2) G_l(p) = \lambda_l \int_0^b p' G_l(p') dp', \quad (116)$$

with
$$\lambda_l = \frac{4\pi\mu}{\hbar^2} (-2)^l a_{2l} \xi^{2l} \gamma_l. \quad (117)$$

In equation (116) we introduce, in order to ensure the convergence of the integral of the second member, an upper limit b to the momentum: this turns out to be equivalent to the introduction of a lower limit for distance in co-ordinate space.*

We obtain in this way

$$G_l(p) = \frac{p^l}{p^2 + p_0^2}, \quad (118)$$

and the energy levels are determined by the condition $1/\lambda_l = T_l(\beta)$, where $\beta = b/p_0$, and

$$T_l(\beta) = p_0^{2l-1} \int_0^\beta \frac{x^{2l} dx}{1+x^2} = p_0^{2l-1} [R_{2l-1}(\beta) - \tan^{-1} \beta], \quad (119)$$

$R_{2l-1}(\beta)$ being a polynomial of degree $2l-1$ in β .

Let us consider now the approximation $(l+1)$

$$K_l^{(l+1)}(p, p') = 2\pi \int_{-1}^{+1} (p^2 + p'^2 - 2pp'x)^{l+1} P_l(x) dx = 4\pi\gamma_l (-2pp')^l (p^2 + p'^2). \quad (120)$$

Consequently, the Schrödinger equation for $G_{l+1}(p)$ can be written

$$\frac{p^2 + p_0^2}{p^l} G_{l+1}(p) - \lambda_l \int_0^b G_{l+1}(p') p' dp' = f_{l+1}(p), \quad (121)$$

with
$$f_{l+1}(p) = 2\lambda_l \frac{a_{2l+2}}{a_{2l}} [T_{l+1}(\beta) + p^2 T_l(\beta)]. \quad (122)$$

Accordingly, G_{l+1} takes the form

$$G_{l+1} = \frac{A_0 + A_1 p^2}{p^2 + p_0^2} p^l, \quad (123)$$

A_0 and A_1 being determined by the equations

$$\left. \begin{aligned} [1 - \lambda_l T_l(\beta)] A_0 &= 2\lambda_l \frac{a_{2l+2}}{a_{2l}} T_{l+1}(\beta), \\ [1 - \lambda_l T_{l+1}(\beta)] A_1 &= 2\lambda_l \frac{a_{2l+2}}{a_{2l}} T_l(\beta). \end{aligned} \right\} \quad (124)$$

* This implies that the method of this section is best fitted to non-relativistic cases: for example, nuclear low energy problems with phenomenological potentials, or molecular problems.

One can continue to calculate in this way the successive approximations of the function G . In a general sense, G_{l+m} will have the form

$$G_{l+m} = \frac{p' S_m(p^2)}{p^2 + p_0^2}, \quad (125)$$

where S_m is a polynomial of degree m in p^2 .

(2) We turn now to the case in which $U(z)$ contains odd powers of z . The question is then rather more complicated. The method, of which we give an outline below, is only practically useful when the parameter ξ is sufficiently small to allow the calculation of only a few terms of the expansion of G . Let us consider the series

$$U(z) = a_1 \xi z + a_3 \xi^3 z^3 + \dots \quad (126)$$

The nucleus $K_l(p, p')$ is expanded in the same way as in the last section

$$K_l(p, p') = \sum_r a_{2r+1} \xi^{2r+1} K_l^{(2r+1)}(p, p'), \quad (127)$$

with
$$K_l^{(2r+1)}(p, p') = 2\pi \int_{-1}^{+1} (p^2 + p'^2 - 2pp'x)^{r+\frac{1}{2}} P_l(x) dx. \quad (128)$$

The approximation (1) corresponds to $r = 0$. In this case we have

$$K_l^{(1)}(p, p') = 2\pi \int_{-1}^{+1} (p^2 + p'^2 - 2pp'x)^{\frac{1}{2}} P_l(x) dx, \quad (129)$$

and, bringing this value into the Schrödinger equation,

$$(p^2 + p_0^2) G_1(p) = \frac{2\pi\mu}{\hbar^{\frac{1}{2}}} \int_0^{+\infty} \int_{-1}^{+1} (p^2 + p'^2 - 2pp'x)^{\frac{1}{2}} G_1(p') P_l(x) p'^2 dp' dx. \quad (130)$$

We transform the integral of the right-hand side according to the general method of § III. We get, in this way,

$$\frac{1}{\cos \frac{1}{2}\psi} G_1(\psi) = \frac{\mu}{\hbar^{\frac{1}{2}}} \int_0^\pi \int_0^\pi \int_0^{2\pi} \frac{1}{\cos^7 \frac{1}{2}\psi'} \sin \frac{1}{2}\Theta G_1(\psi') P_l(\cos \theta') d\Omega'. \quad (131)$$

Let us put
$$G_1(\psi) = \sum_n a_n \cos^7 \frac{1}{2}\psi P_{n,l}^{(2)}(\cos \psi). \quad (132)$$

The application of the theorem of Hecke gives the relation

$$\sum_n \left[(1 + \cos \psi)^3 - \frac{\mu}{\hbar^{\frac{1}{2}}} \Lambda_n \right] a_n P_{n,l}^{(2)}(\cos \psi) = 0, \quad (133)$$

with
$$\Lambda_n = \frac{2^{\frac{1}{2}}\pi}{n+1} \int_{-1}^{+1} (1-x)^{\frac{1}{2}} C_n^1(x) (1-x^2)^{\frac{1}{2}} dx. \quad (134)$$

The repeated application of the recurrence formula (A 7) transforms the expression under the summation index in equation (133) in a sum of seven functions:

$$P_{n-3,l}^{(2)}, P_{n-2,l}^{(2)}, \dots, P_{n+3,l}^{(2)}.$$

We recombine these terms with the other terms of the series, and we equate separately to zero the coefficients of $P_{n,l}^{(2)}$ (which are linear expressions in $a_{n-3}, a_{n-2}, \dots, a_{n+3}$).

In this way we arrive at a system of linear equations which allow a successive determination of the a_n . The energy levels equation is expressed, in the general case, as a compatibility relation between these equations. As it has been seen in §IV (relativistic hydrogen atom) this relation is sometimes very easy to find.

VII. CONCLUSION

The author hopes to have demonstrated the interest and the various possibilities of the momentum representation for the study of problems where the interaction potential is a central one, or, more generally, can be expressed in this representation in a simple form. The main difficulty which is encountered is, of course, the solution of integral equations which appear in this representation. The general methods of solution (the successive approximations method of Neumann-Liouville, or the variation-iteration method of Svartholm) can obviously be applied to these equations. But emphasized in this paper are those methods which are particularly fitted to the types of integral equations which correspond, in momentum space, to Schrödinger and Dirac equations. In the case of the Schrödinger equation, for example, the method of expansion of the kernel in a series of functions, used in §§V and VI, can give satisfactory results. As for the Dirac system of integral equations, the various procedures which have been mentioned in order to reduce the problem to the solution of one or two Schrödinger equations (reduction of the system or successive approximations method based on the separation between large and small components) possesses a large field of application. In both cases a unitary form has been given to the different solution procedures by means of the powerful method of passage to Fock's hyperspace; this method allows one, by an appropriate use of Hecke's theorem and of the recurrence formulae, to express the solution as a linear combination of hyperspherical harmonics. Moreover, the fact that a rigorous solution of the Coulombian problem is known, can determine sometimes the choice of the best successive approximations method, as it is the case, for example, for the 'meson' potential.

The author would like to express his gratitude to Professor L. Rosenfeld for the interest which he has shown in this work and for many illuminating discussions. He is also indebted to Professor R. E. Peierls for his helpful suggestions in the presentation of the paper.

APPENDIX. MATHEMATICAL RESULTS USED IN THE TEXT

(1) *Integral equations of hyperspherical harmonics (Hecke 1918)*

For the construction and general properties of hyperspherical harmonics, the reader should refer to Appell & Kampé de Fériet, (1926) and Magnus & Oberhettinger (1948).

Let us consider in a $(q + 2)$ -dimensional space, two unit vectors $\mathbf{u}$ and $\mathbf{v}$, expressed, in polar co-ordinates, respectively by means of $\theta_1, \theta_2, \dots, \theta_q, \phi$ and $\theta'_1, \theta'_2, \dots, \theta'_q, \phi'$.

Let $F(\cos \Theta)$ be a function depending only on the product $\mathbf{u} \cdot \mathbf{v}$. The general integral equation of hyperspherical harmonics is the following:

$$\phi(\mathbf{u}) = \lambda_n \int_{\omega_{q+1}(\mathbf{v})} F(\cos \Theta) \phi(\mathbf{v}) d\omega_{(q+1)}(\mathbf{v}), \quad (\text{A } 1)$$

with
$$\frac{1}{\lambda_n} = \frac{2\pi^{\frac{1}{2}(q+1)}}{\Gamma[\frac{1}{2}(q+1)]} \int_{-1}^{+1} F(x) P_{n,0}^{(q)}(x) (1-x^2)^{\frac{1}{2}(q-1)} dx, \quad (\text{A } 2)$$

where we have put
$$P_{n,0}^{(q)}(x) = \frac{n!(q-1)!}{(n+q-1)!} C_n^{\frac{1}{2}q}(x). \quad (\text{A } 3)$$

When the integral (A 2) does not exist, one can generalize by means of a method given by Erdelyi (1938), to which the reader should refer for further details.

(2) Properties of Legendre functions used in the text

(a) Passage from the functions of $\cos 2\theta$ to the functions of $\cos \theta$

We start from the following integral representation of Legendre functions (Hobson 1931, p. 248):

$$P_\nu^{-\mu}(\cos 2\theta) = \frac{e^{\frac{1}{2}(i\pi\mu)} \sin^\mu 2\theta}{2^\mu \Gamma(\frac{1}{2}) \Gamma(\mu + \frac{1}{2})} \int_0^\pi (\cos 2\theta + i \sin 2\theta \cos \psi)^{\nu-\mu} \sin^{2\mu} \psi d\psi, \quad (\text{A } 4)$$

which can be written

$$P_\nu^{-\mu}(\cos 2\theta) = (-1)^{\nu-\mu} \frac{e^{\frac{1}{2}(i\pi\mu)} \sin^\mu 2\theta}{2^\mu \Gamma(\frac{1}{2}) \Gamma(\mu + \frac{1}{2})} \int_0^\pi [1 - 2 \cos \theta (\cos \theta + i \sin \theta \cos \psi)]^{\nu-\mu} \sin^{2\mu} \psi d\psi. \quad (\text{A } 5)$$

We expand the expression between brackets under the integration sign of the right-hand side, according to the binomial formula. If $\nu - \mu$ is an integer (which is the case in §IV of the text), we get (taking into account relation (A 4))

$$P_\nu^{-\mu}(\cos 2\theta) = (-1)^{\nu-\mu} 2^\mu \cos^\mu \theta \sum_{k=0}^{\nu-\mu} \frac{(-2)^k \Gamma(\nu-\mu+1)}{k! \Gamma(\nu-\mu-k+1)} \cos^k \theta P_{\mu+k}^{-\mu}(\cos \theta), \quad (\text{A } 6)$$

which is the required relation. If $\nu - \mu$ is not an integer, the finite sum of the right-hand side of (A 6) is replaced by an infinite sum (the summation extends between 0 and $+\infty$).

(b) Recurrence formulae (Magnus & Oberhettinger 1948, p. 81)

$$(2\nu+1)xP_\nu^\mu(x) = (\nu-\mu+1)P_{\nu+1}^\mu(x) + (\nu+\mu)P_{\nu-1}^\mu(x), \quad (\text{A } 7)$$

$$P_{\nu-1}^\mu(x) - P_{\nu+1}^\mu(x) = (2\nu+1)\sqrt{(1-x^2)}P_\nu^{\mu-1}(x), \quad (\text{A } 8)$$

$$(\nu-\mu)(\nu-\mu+1)P_{\nu+1}^\mu(x) - (\nu+\mu)(\nu+\mu+1)P_{\nu-1}^\mu(x) = (2\nu+1)\sqrt{(1-x^2)}P_\nu^{\mu+1}(x). \quad (\text{A } 9)$$

(3) Fourier transformation of a central potential of the type r^ν (ν positive or negative)

Let us consider in a $(q+2)$ -dimensional space the Fourier transform of a central potential equal to a power of r or $1/r$. In the calculation of this transform the introduction of a convergence factor of the type e^{-kr} is generally needed. This leads

to investigating, in the general case, a potential of the kind $r^\nu e^{-kr}$ (which, in the case $\nu = -1$, is identical with the Yukawa potential). One can, afterwards, make $k \rightarrow 0$ if one wishes.

The transform of such a potential is defined by

$$U_k^{(\nu)}(\mathbf{p}) = \frac{1}{h^{\frac{1}{2}q+1}} \iint \dots \int_{-\infty}^{+\infty} r^\nu e^{-kr} e^{i(p_1 x_1 + p_2 x_2 + \dots + p_{q+2} x_{q+2})/\hbar} dx_1 dx_2 \dots dx_{q+2}. \quad (\text{A } 10)$$

In polar co-ordinates we put

$$d\omega_{q+2} = r^{q+1} dr \sin^q \theta_1 \sin^{q-1} \theta_2 \dots \sin \theta_q d\theta_1 \dots d\theta_q d\phi, \quad (\text{A } 11)$$

and have to calculate

$$U_k^{(\nu)}(\mathbf{p}) = \frac{1}{h^{\frac{1}{2}q+1}} \int_{r=0}^{+\infty} \int_{\theta_1=0}^{\pi} \dots \int_{\theta_q=0}^{\pi} \int_{\phi=0}^{2\pi} r^\nu e^{-kr} e^{ir(p_1 u_1 + p_2 u_2 + \dots + p_{q+2} u_{q+2})/\hbar} d\omega_{q+2}. \quad (\text{A } 12)$$

$u_1, \dots, u_{q+2}$ are the components of vector $\mathbf{u}$ introduced in the last section. In $U_k^{(\nu)}(\mathbf{p})$, these components appear in a symmetrical way. Let us put $p = (\sum_r p_r^2)^{\frac{1}{2}}$; then we are entitled to write $U_k^{(\nu)}(\mathbf{p})$ in the form

$$U_k^{(\nu)}(p) = \frac{1}{h^{\frac{1}{2}q+1}} \int_{r=0}^{+\infty} \int_{\theta_1=0}^{\pi} \dots \int_{\theta_q=0}^{\pi} \int_{\phi=0}^{2\pi} r^\nu \exp \left[-kr + \frac{ipr}{\hbar} \cos \theta_1 \right] d\omega_{q+2}, \quad (\text{A } 13)$$

and, using Poisson's formula (Watson 1922, p. 48);

$$U_k^{(\nu)}(p) = \frac{(2\pi)^{-\nu} h^{\nu+\frac{1}{2}q+1}}{2p^{\nu+q+2}} \int_0^{+\infty} e^{-k\hbar z/p} J_{\frac{1}{2}q}(z) z^{\nu+\frac{1}{2}q+1} dz. \quad (\text{A } 14)$$

The integral on the right-hand side is convergent for $z \rightarrow 0$ only when $\nu + q + 2 > 0$. In this case we put $p/k\hbar = \tan \theta$, and we obtain, using the Laplace transform of the product $z^{\nu+\frac{1}{2}q+1} J_{\frac{1}{2}q}(z)$,

$$U_k^{(\nu)}(\theta) = \frac{\left(\frac{2\pi}{\hbar}\right)^{\frac{1}{2}q+1}}{2k^{\nu+q+2}} \Gamma(\nu + q + 2) \cos^{\nu+q+2} \theta \sin^{-\frac{1}{2}q} \theta P_{\nu+\frac{1}{2}q+1}^{-\frac{1}{2}q}(\cos \theta). \quad (\text{A } 15)$$

If k tends to zero, $\cos \theta/k \rightarrow \hbar/p$, and

$$P_{\nu+\frac{1}{2}q+1}^{-\frac{1}{2}q}(\cos \theta) = \frac{\sqrt{\pi} 2^{-\frac{1}{2}q}}{\Gamma\left(\frac{\nu+q+3}{2}\right) \Gamma\left(-\frac{\nu}{2}\right)}. \quad (\text{A } 16)$$

Consequently

$$U_0^{(\nu)}(p) = \frac{\pi^{\frac{1}{2}(q+3)} \Gamma(\nu + q + 2)}{\Gamma\left(\frac{\nu+q+3}{2}\right) \Gamma\left(-\frac{\nu}{2}\right)} \frac{\hbar^{\nu+\frac{1}{2}q+1}}{p^{\nu+q+2}}. \quad (\text{A } 17)$$

In particular, one can write the following correspondences:

$$\frac{1}{r} \rightarrow \frac{\pi^{\frac{1}{2}q+1} \Gamma(q+1)}{\Gamma(\frac{1}{2}q+1)} \frac{\hbar^{\frac{1}{2}q}}{p^{q+1}} \quad (q+2 > 1), \quad (\text{A } 18)$$

$$\frac{1}{r^2} \rightarrow \frac{\pi^{\frac{1}{2}(q+3)} \Gamma(q)}{\Gamma(q+\frac{1}{2})} \frac{\hbar^{\frac{1}{2}q-1}}{p^q} \quad (q+2 > 2), \quad (\text{A } 19)$$

$$\frac{1}{r^3} \rightarrow \frac{2\pi^{\frac{1}{2}q+1} \Gamma(q-1)}{\Gamma(\frac{1}{2}q)} \frac{\hbar^{\frac{1}{2}q-2}}{p^{q-1}} \quad (q+2 > 3). \quad (\text{A } 20)$$

It can be seen that the correspondence in momentum space of potentials of the type $1/r_n$ (and the resultant divergence) depend essentially on the number of dimensions of the space where the transformation is effected. This is of importance for the study of relativistic interactions between particles, where four- or five-dimensional spaces are commonly used.

Case of the Yukawa potential. If one makes $\nu = -1$ in equation (A 15), one gets

$$U_k^{(-1)}(\theta) = \frac{\left(\frac{2\pi}{\hbar}\right)^{\frac{1}{2}q+1}}{2k^{q+1}} \Gamma(q+1) \cos^{q+1} \theta \sin^{-\frac{1}{2}q} \theta P_{\frac{1}{2}q}^{-\frac{1}{2}q}(\cos \theta). \quad (\text{A } 21)$$

The Legendre function which appears on the right-hand side has the degenerate value

$$P_{\frac{1}{2}q}^{-\frac{1}{2}q}(\cos \theta) = \frac{1}{\Gamma(\frac{1}{2}q+1)} \left(\frac{1}{2} \sin \theta\right)^{\frac{1}{2}q}. \quad (\text{A } 22)$$

Consequently

$$U_k^{(-1)}(\theta) = \left(\frac{\pi}{\hbar}\right)^{\frac{1}{2}q+1} \frac{\Gamma(q+1)}{\Gamma(\frac{1}{2}q+1)} \left(\frac{\cos \theta}{k}\right)^{q+1}, \quad (\text{A } 23)$$

and, coming back to the variable p ,

$$U_k^{(-1)}(\theta) = \pi \hbar^{\frac{1}{2}q} \frac{\Gamma(q+1)}{\Gamma(\frac{1}{2}q+1)} \frac{1}{(p^2 + k^2 \hbar^2)^{\frac{1}{2}(q+1)}}. \quad (\text{A } 24)$$

REFERENCES

- Appell, P. & Kampé de Fériet, J. 1926 *Fonctions hypergéométriques et hypersphériques*. Paris: Gauthier-Villars.
- Bethe, H. 1933 *Handb. der Phys.* 2nd ed., p. 273. Berlin: Springer.
- Dirac, P. A. M. 1927 *Z. Phys.* **44**, 585.
- Dirac, P. A. M. 1935 *Ann. Math.* **36**, 657.
- Elsasser, W. 1933 *Z. Phys.* **81**, 332.
- Erdelyi, A. 1938 *Math. Ann.* **115**, 456.
- Fock, V. 1935 *Z. Phys.* **98**, 145.
- Gordon, W. 1934 *Skand. Mat. Kongr. (Stockholm)*, p. 249.
- Hecke, E. 1918 *Math. Ann.* **78**, 398.
- Hobson, E. W. 1931 *The theory of spherical and ellipsoidal harmonics*. Cambridge University Press.
- Hylleraas, E. A. 1932 *Z. Phys.* **74**, 216.
- Lichnerowicz, A. 1947 *Algèbre et analyse linéaires*. Paris: Masson et Cie.
- Magnus, W. & Oberhettinger, F. 1948 *Formeln und Sätze für die speziellen Funktionen der Mathematischen Physik*, 2nd ed. Berlin: Springer.
- Mott, N. F. & Sneddon, I. N. 1948 *Wave mechanics and its applications*. Oxford University Press.
- Podolsky, B. & Pauling, L. 1929 *Phys. Rev.* **34**, 109.
- Rubinowicz, A. 1948 *Phys. Rev.* **73**, 1330.
- Svartholm, N. 1945 The binding energies of the lightest atomic energies. Thesis, Lund.
- Van Hove, L. 1949 *Phys. Rev.* **75**, 1519.
- Watson, G. N. 1922 *Theory of Bessel functions*, 1st ed. Cambridge University Press.
- Weyl, H. 1928 *Z. Phys.* **48**, 1.
- Whittaker, E. T. & Watson, G. N. 1927 *A course of modern analysis*, 4th ed. Cambridge University Press.

An interferometer microscope

By J. DYSON

Research Laboratory, Associated Electrical Industries Ltd., Aldermaston, Berks

*(Communicated by T. E. Allibone, F.R.S.—Received 25 May 1950—
Revised 21 July 1950)*

[Plates 4 to 7]

Some of the drawbacks of conventional phase contrast microscopy are enumerated and the desirability of an interferometer microscope is indicated. A number of existing ways of achieving this are described, together with their merits and drawbacks. A new type of interferometer microscope which avoids some of these disadvantages is described, together with its operational procedure. The conditions covering coherence in the object plane are investigated and a simple approximation for evaluating the coherence is derived. An elementary discussion of contrast arising from phase or amplitude objects is given, and a number of experimental results are presented as photo-micrograms.

Finally, a method for the accurate measurement of the optical thickness of an object is described, with the results of a determination of the refractive index of the cytoplasm of epithelial cells.

INTRODUCTION

It is well known that the conventional methods of phase contrast microscopy have certain features which are undesirable for some types of work. One such feature arises from the fact that contrast is obtained by causing interference between the illuminating beam and light which has been diffracted outside this beam by the object and thus does not pass through the phase ring. As the light falling outside the phase ring has for the greater part been diffracted through appreciable angles (of the order of the angular width of the phase ring subtended at the object), it follows that the bulk of this light has been diffracted by fine detail and sharp edges in the object.

This fact is reflected in the characteristic appearance of phase contrast images. There is a strong tendency for boundaries between areas of differing optical thicknesses to be delineated by an abrupt change of intensity bordered by light and dark fringes with the intensities at large distances from the boundary tending to become equal. This is shown in figure 1. A distribution of optical thickness of the form illustrated in figure 1*a* is represented by an intensity distribution as shown in figure 1*b*. This phenomenon is known as the 'halo effect'. Thus the edge of the image of a cell, for example, is bordered by a bright fringe on the outside. If the cell be of uniform thickness and sufficiently large, a dark fringe will appear just inside the edge; this fringe may be reduced by the effect of non-uniformity of thickness but, conversely, its presence causes difficulty in appreciation of variations of thickness across the cell.

For many applications this 'optical artefact' is of small importance, but the presence of detail in the image which is not present in the object is obviously undesirable.

Further, it is not possible with conventional systems to vary continuously the amount of retardation introduced between the two interfering beams. This may be done by methods involving the use of polarized light, but, as many objects suitable for examination by phase contrast are birefringent, effects other than straight-forward phase contrast may be introduced.

In addition, the absence of means for variation of retardation causes difficulty in the measurement of optical thickness of the object.

These considerations led the author to seek for a method which would reveal changes of optical thickness and which was free from the above objections.

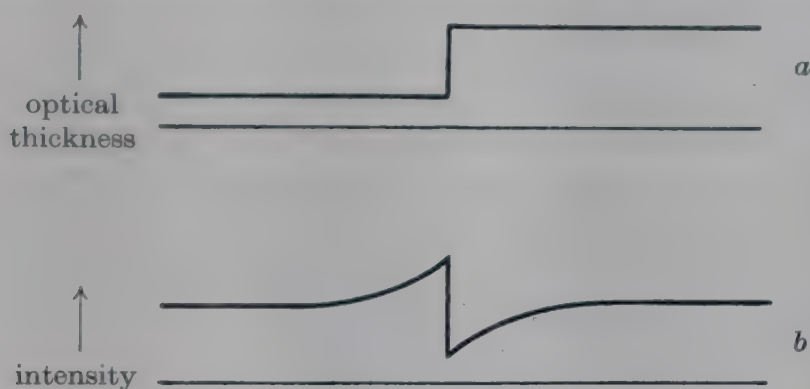


FIGURE 1. Intensity distribution near the image of an abrupt change of optical thickness in conventional phase contrast microscopy.

INTERFERENCE MICROSCOPY

A new method of showing variations in optical thickness has been devised independently by Frederikse (1933) and Merton (1947*a, b*). The object is placed between two half-silvered glass plates and illuminated by transmission, thereby forming in effect a Fabry-Perot interferometer with the object between the plates. If the plate separation is such as to give extinction by interference of the light traversing the plates, the field will appear dark, but the presence of the object will upset the condition for destructive interference of the light passing through it, and hence the object will appear bright on a dark ground.

This simple description applies, however, only to the case of parallel illumination of the object, and even then is only on approximation, neglecting the effects of diffraction at the object. It may readily be seen that the effect of the two half-silvered plates is to produce an infinite number of virtual images of the object, equally spaced along the common normal to the two plates. The object itself is seen by light which has passed through it once only; the virtual images by light which has traversed it 3, 5, 7, ..., $(2n + 1)$... times respectively.

If all these images could be made to interfere with each other simultaneously, the very high sensitivity to differences of optical path characteristic of a Fabry-Perot interferometer or of Tolansky's multiple-beam techniques would be realized. However, the images are separated by twice the separation of the two plates, and, as this amounts to several microns at the least, it is not possible to bring more than one image at a time into focus if an appreciable illuminating aperture is used.

Merton (1947*b*) has overcome part of this difficulty by placing a zone plate in the back focal plane of the condenser. This causes a number of equidistant real images of the light source to appear in the object space, and may be so arranged that one such image illuminates the object while another coincides with one of the virtual images of the object. If the microscope be focused on this virtual image two coincident fields of view are seen: one contains the image of the light source only, the other the image of the light source with the object superimposed. There is a path-difference of twice the plate separation between these images, but if sufficiently monochromatic light be used interference can take place between them, giving the desired representation of difference of optical thickness as variations of intensity with a comparatively large illuminating cone.

This system, however, still suffers from some disadvantages. The path-difference cannot be made very small, so white light cannot be used; furthermore, the path-difference cannot easily be varied without disturbing the specimen. Also, the construction of a zone plate for a large illuminating aperture is somewhat difficult, and the technique depends for its success on the perfection of the condenser. Against these considerations is the fact that the method may be used on a conventional microscope stand, practically without modification.

TWO-OBJECTIVE INTERFEROMETER MICROSCOPE

These optical disadvantages can be overcome, at the expense of introducing mechanical difficulties, by a type of interferometer shown diagrammatically in figure 2. This is a 'round-the-square' interferometer in which the light beam is split before entering the microscope system. A condenser-objective system is placed in each beam, one such system containing the object, and the two beams are recombined by a half-silvered mirror between objective and eyepiece.

This system enables an unrestricted aperture of illumination to be used, but the two objectives and two condensers must be carefully matched to ensure identical

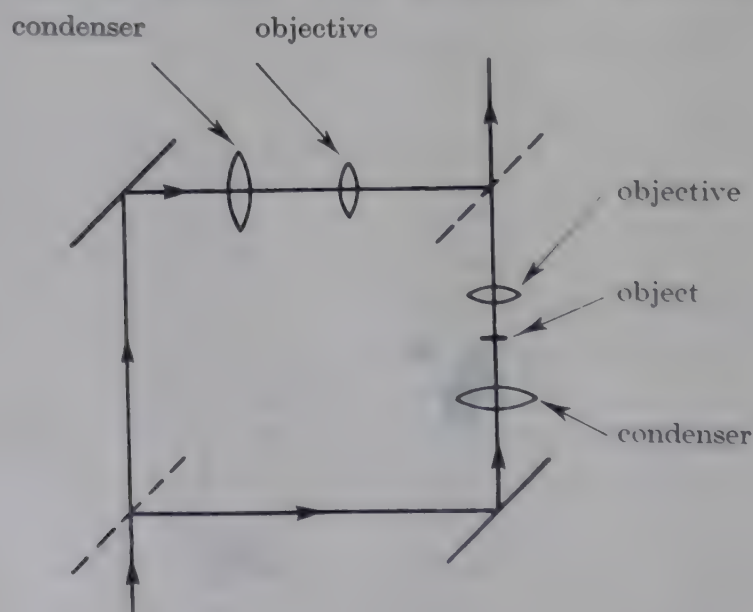


FIGURE 2. Two-objective interferometer microscope.

magnification, aberrations and chromatic difference of path length if the use of white light is desired. A more serious difficulty is the necessary mechanical rigidity of the supporting stand and perfection of the various adjustments, with consequent high cost of the apparatus. A modification of this system for metallurgical applications has been designed by Linnik and manufactured by Messrs Zeiss.

OTHER TYPES OF INTERFEROMETER

Various other types of instrument suitable for microscopy have been designed by Bates (1947), Philpot, Hopkins and others, but in most cases these have not been described in published work. Some of these instruments, however, require some restrictions on the nature of the illumination; for example, in the case of Bates's 'wave-front shearing' interferometer the aperture of illumination is restricted to a narrow slit.

A NEW TYPE OF INTERFEROMETER MICROSCOPE

The mechanical problems can be considerably simplified if the division and recombination of the beams is effected in the space between condenser and objective. A method of doing this is shown in figure 3. The illuminating cone from the condenser is intercepted by a nearly plane-parallel glass plate, the top surface of which is half-silvered, while the bottom surface bears a small, fully silvered central spot somewhat larger than the field of view of the microscope. Part of the illuminating cone traverses the plate and illuminates the object, whereas part is reflected downwards and reflected upwards once more by the silvered spot. The plate is so disposed that the light source is approximately focused on the silvered spot, so the whole of this portion of the beam is intercepted thereby.

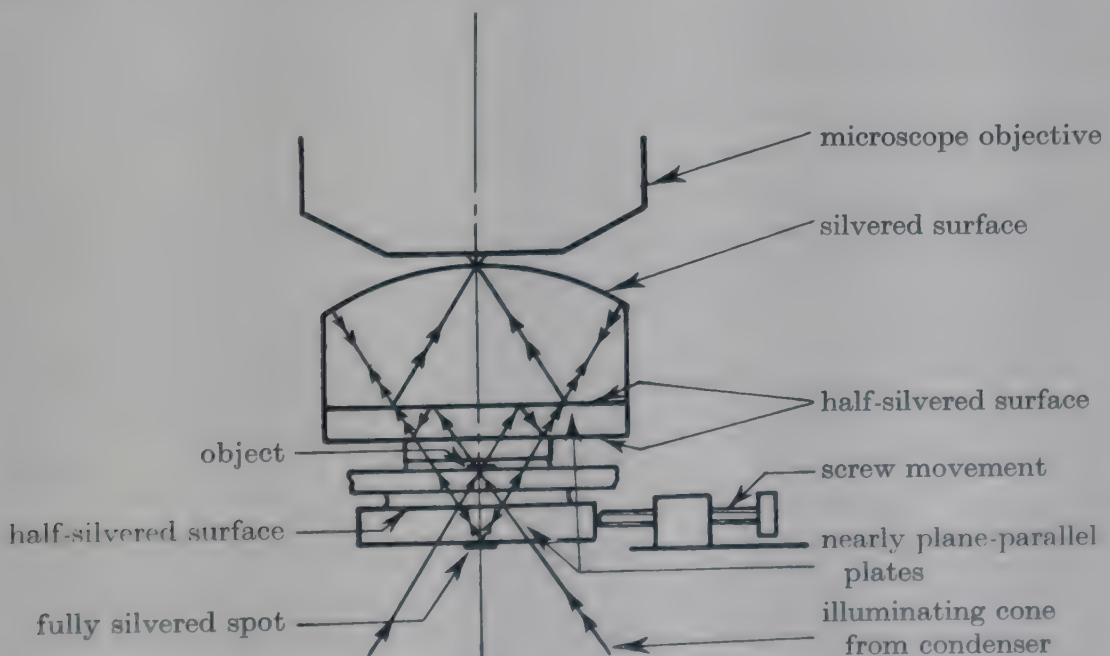


FIGURE 3. Optical system of interferometer microscope.

The plane of the object is thus traversed by two beams. Of these that which suffered two internal reflexions in the plate and which may be called the 'comparison beam' has a diameter of some millimetres in the object plane, and so the presence of

an object comparable in size with the microscope field does not disturb it and will, in fact, lie within the central 'missing cone' which is stopped out by the silvered spot on the first entry of the light into the plate.

The two beams are recombined by means of a second plate, identical with the first except that it is half-silvered on both faces, placed above the object. Part of the comparison beam traverses this plate without reflexions, and part of the illuminating beam which has passed through the object suffers two internal reflexions, finally emerging from the upper surface. If the two plates are correctly aligned, the two beams will coincide. An observer situated above the upper plate will see two coincident fields of view; one of these contains the image of the light source only, the other the image of the light source with the object superimposed. The path-difference between these images, under the conditions described, is zero except for any variations in the optical thickness of the object, or any small shift, uniform over the field, due to unequal phase-shifts caused by reflexion at the silvered surfaces. Conditions are thus suitable for interference to take place, in white light if desired, and the necessary representation of optical thickness in terms of intensity will be exhibited. The two images are situated in the vicinity of the silvered spot on the lower plate and are therefore beyond the reach of a conventional objective of high power. To overcome this difficulty, a method described in a previous paper (Dyson 1949) is employed. A glass block with a spherical upper surface is cemented to the upper plate. The centre of curvature of the block coincides with the axial point of the images, and the radius is such that it is bisected by the top surface of the upper plate. The spherical surface is silvered except for a small spot on the axis.

Under these conditions a real image is formed, after reflexion at the hemispherical surface and at the half-silvered surface of the top plate, in approximate coincidence with the spherical surface of the glass block. The light can emerge into air through the hole in the silvering and the image may then be observed by means of a conventional microscope objective. No aberrations which are serious for the fields of view concerned are introduced by this system.

It will be noted that a large number of reflexions other than those mentioned above will take place in the parallel-plate system. However, all the images produced by these reflexions are at least several millimetres out of focus. The out-of-focus disks produced thereby have black centres, on account of the stopping out of the centre of the illuminating cone by the silvered spot on the lower plate, and the disks are so large that the black centre is in all cases larger than the field of the microscope. Hence, these unwanted reflexions do not decrease the contrast of the image, but merely lead to a loss of light.

If the plates be made not plane-parallel but with a slight wedge angle a very convenient method of varying the path-difference between the two interfering fields may be obtained. If the wedge angle be θ and one of the plates be translated in a direction normal to the wedge axis and to the optical axis through a distance d , the effective thickness of the plate is changed by an amount θd . The path difference between the two fields is thus changed by an amount $2\theta d$. If, for example, $\theta = \frac{1}{400}$, then to change the path-difference by one wave-length the plate must be translated by 200λ , where λ is the wave-length in glass. This corresponds to a movement of

about 0.1 mm., which is very easily achieved by means of a screw adjustment without any particular refinements.

It is usually mechanically easier to arrange to translate the lower plate, as the upper one is fixed to the microscope objective. This means that the fully silvered spot is also translated, and so the wedge angle must be selected sufficiently large to ensure that the desired path-difference can be achieved before the silvered spot moves too far from the central position.

To ensure that the path-difference is uniform over the field the wedge axes of the upper and lower plates must be parallel. If they make an angle ϕ with each other, the path-difference will vary linearly across the field, which will be crossed by parallel equidistant fringes of separation equal to $\lambda/2\theta(1 - \cos \phi)$.

For this reason adjustments for both translation and rotation of at least one plate must be provided. In addition, means for tilting one plate must be available, for the effect of tilt is to translate one image laterally with respect to the other. It is evident that interference will only take place when the two images of the light source are so superimposed that the images of corresponding points in the light source coincide, for the light from widely separated points will not be coherent. The accuracy with which this must be effected will now be investigated.

COHERENCE IN THE OBJECT PLANE

If the light incident at a given point in the plane of the object be considered, then it is evident that, in general, it will be only partially coherent with the light incident at some other point. Furthermore, it will usually be true that the degree of coherence will decrease as the separation of the two points increases.

The coherence may be evaluated by means of a hypothetical experiment. Suppose that an opaque screen pierced by two very small holes separated by a distance d be placed in the object plane. The diameters of the holes are so chosen that the total intensity of light passing through each is equal. The two beams of light so obtained are allowed to fall on another screen so disposed that a path-difference p , which is a function of position on this screen, is introduced between the two beams. Then a series of bright and dark fringes is formed, in which the maximum and minimum intensities are $I_{\max.}$ and $I_{\min.}$ respectively.

Following Zernike (1938, 1948) and others, we take as the measure of the coherence between the light emerging through the two holes the 'visibility' V of these fringes, defined as

$$V = \frac{I_{\max.} - I_{\min.}}{I_{\max.} + I_{\min.}}.$$

Thus, when the coherence is complete, the dark fringes being perfectly black, $I_{\min.} = 0$ and $V = 1$. Similarly, when interference cannot take place, $I_{\max.} = I_{\min.}$ and $V = 0$.

The value of V has been obtained (Van Cittert 1934) for two cases, viz. that of illumination by a uniformly bright extended source with no intervening optical system, and that of a uniformly bright source sharply imaged on to the plane by a perfect condenser lens. Van Cittert showed that in either case the value of V for

a separation d was equal to the amplitude in the diffraction pattern of a perfect lens, the aperture of which subtended the same solid angle at the screen as either the source or the aperture of the condenser lens, taking the amplitude of the central maximum of the diffraction pattern as unity. The case of an imperfect condenser which is filled with light has been considered by Zernike (1938), but the case of a condenser which is not filled with light does not seem to have been considered.

Consider an extended light source (figure 4) separated from the object plane by some condenser system about which we assume only that it is free from absorption. The object plane is pierced by two very small holes A and B of equal size and distant d apart.

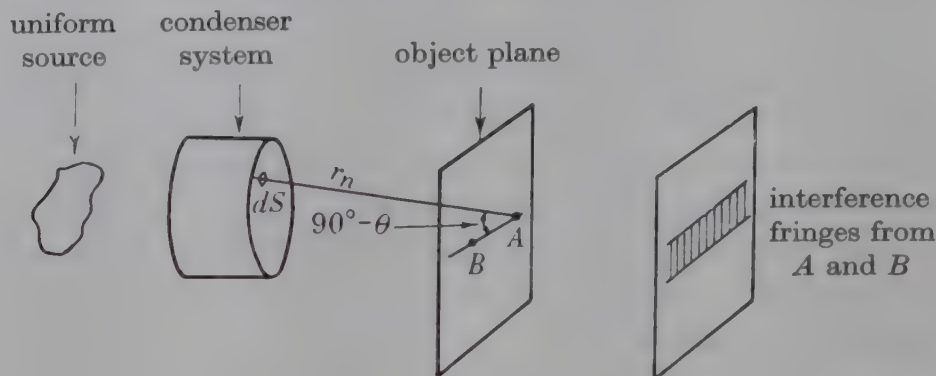


FIGURE 4. Coherence in the object plane.

We make the following assumptions about the light source:

- (a) It is of uniform brightness.
- (b) It consists of a large number of independently radiating atoms.
- (c) The light received at any point from one such atom is capable of interfering with the light received from the same atom at any other point.

Of these assumptions (a) depends on the experimental arrangement and is the ideal aimed at in practice. (b) is true of a gas discharge and is almost certainly sufficiently true of an incandescent solid. (c) is true if each atom behaves as a single radiator of spherical waves.

A further assumption is that either the source is so large, or that it is focused so far away from the object plane, or that the condenser is so imperfect, that the intensities of illumination at A and B are sensibly equal. The illumination on the object plane is then due to a complex of wave-fronts, one from each radiating atom, emerging from the condenser. We represent the vector amplitudes due to the n th wave-front at A and B as $\tilde{E}_{no}, \tilde{E}_{nd}$, and the magnitudes of these vectors as E_{no}, E_{nd} . The light passing through A and B is recombined on a further screen, with a phase-difference p which is a function of position on the screen, to form interference fringes. The amplitude on this screen due to the combination of $\tilde{E}_{no}, \tilde{E}_{nd}$ is represented by E_n . Then

$$E_n^2 = E_{no}^2 + E_{nd}^2 + 2E_{no}E_{nd}\cos(p + \theta_n), \quad (1)$$

where the path-difference p is measured in radians and where θ_n is the phase-difference between $\tilde{E}_{no}$ and $\tilde{E}_{nd}$.

The illumination I on the screen is equal to $\sum_n E_n^2$, for, as wave-fronts with different n are independent, the illumination due to each must be added.

Therefore

$$I = \sum_n E_{no}^2 + \sum_n E_{nd}^2 + 2 \sum_n E_{no} E_{nd} \cos(p + \theta_n). \quad (2)$$

$$\text{Now } \sum_n E_{no} E_{nd} \cos(p + \theta_n) = [\sum_n E_{no} E_{nd} \cos \theta_n] \cos p - [\sum_n E_{no} E_{nd} \sin \theta_n] \sin p.$$

This is a quantity sinusoidal in p , with a maximum value M given by

$$M^2 = [\sum_n E_{no} E_{nd} \cos \theta_n]^2 + [\sum_n E_{no} E_{nd} \sin \theta_n]^2. \quad (3)$$

As a result

$$I_{\max.} = \sum_n E_{no}^2 + \sum_n E_{nd}^2 + 2M$$

and

$$I_{\min.} = \sum_n E_{no}^2 + \sum_n E_{nd}^2 - 2M,$$

giving

$$V = \frac{I_{\max.} - I_{\min.}}{I_{\max.} + I_{\min.}} = \frac{2M}{\sum_n E_{no}^2 + \sum_n E_{nd}^2}. \quad (4)$$

The denominator of this expression is easily interpreted, for it represents simply the sum of the intensities of illumination at A and B . The conditions have been chosen such that these intensities are sensibly uniform, so the denominator may be equated to twice the illumination I_0 in the object plane.

Thus

$$V = \frac{M}{I_0}. \quad (5)$$

To evaluate M , we take the amplitude of the n th wave-front as it leaves the condenser aperture S as being e_n . The length of a straight line from A drawn through an element dS of the condenser aperture to meet the n th wave-front is taken as r_n , being a function of the position of dS . This line makes an angle θ with the plane to which the line joining A and B is normal.

Then

$$\tilde{E}_{no} = \int_S e_n \exp\left\{\frac{i2\pi r_n}{\lambda}\right\} dS \quad (6)$$

and

$$\begin{aligned} \tilde{E}_{nd} &= \int_S e_n \exp\left\{\frac{i2\pi}{\lambda}(r_n + d \sin \theta)\right\} dS \\ &= \int_S \exp\left\{\frac{i2\pi d \sin \theta}{\lambda}\right\} \left[e_n \exp\left\{\frac{i2\pi r_n}{\lambda}\right\} dS \right], \end{aligned} \quad (7)$$

where λ is the wave-length.

But, $e_n \exp\left\{\frac{i2\pi r_n}{\lambda}\right\} dS = d\tilde{E}_{no}$, the contribution to $\tilde{E}_{no}$ from dS . Therefore (7) becomes

$$\tilde{E}_{nd} = \int_S \exp\left\{\frac{i2\pi d \sin \theta}{\lambda}\right\} d\tilde{E}_{no}. \quad (8)$$

We note that, as $\tilde{E}_{no}, \tilde{E}_{nd}$ are two-dimensional vectors, we can write (3) as

$$M^2 = [\sum_n \tilde{E}_{no} \cdot \tilde{E}_{nd}]^2 + [\sum_n \tilde{E}_{no} \times \tilde{E}_{nd}]^2, \quad (9)$$

interpreting $\tilde{E}_{no} \times \tilde{E}_{nd}$ as the magnitude of the vector product of $\tilde{E}_{no}$ and $\tilde{E}_{nd}$. Then, from (8),

$$M^2 = \left[\sum_n \tilde{E}_{no} \cdot \int_S \exp \left\{ \frac{i2\pi d \sin \theta}{\lambda} \right\} d\tilde{E}_{no} \right]^2 + \left[\sum_n \tilde{E}_{no} \times \int_S \exp \left\{ \frac{i2\pi d \sin \theta}{\lambda} \right\} d\tilde{E}_{no} \right]^2. \quad (10)$$

We write $\tilde{E}_{no}$ as $\tilde{A}_{no}$, as a reminder that it is a constant during the indicated integrations, and take it inside the integrals, remembering that $\exp(i2\pi d \sin \theta / \lambda)$ is an operator rotating the vector $d\tilde{E}_{no}$ through an angle $\pi d \sin \theta / \lambda$. Then (10) becomes

$$\begin{aligned} M^2 &= \left[\sum_n \int \cos \frac{2\pi d \sin \theta}{\lambda} \tilde{A}_n \cdot d\tilde{E}_n - \sum_n \int \sin \frac{2\pi d \sin \theta}{\lambda} \tilde{A}_n \times d\tilde{E}_n \right]^2 \\ &\quad + \left[\sum_n \int \cos \frac{2\pi d \sin \theta}{\lambda} \tilde{A}_n \times d\tilde{E}_n + \sum_n \int \sin \frac{2\pi d \sin \theta}{\lambda} \tilde{A}_n \cdot d\tilde{E}_n \right]^2 \\ &= \left[\int \cos \frac{2\pi d \sin \theta}{\lambda} \sum_n (\tilde{A}_n \cdot d\tilde{E}_n) - \int \sin \frac{2\pi d \sin \theta}{\lambda} \sum_n (\tilde{A}_n \times d\tilde{E}_n) \right]^2 \\ &\quad + \left[\int \cos \frac{2\pi d \sin \theta}{\lambda} \sum_n (\tilde{A}_n \times d\tilde{E}_n) + \int \sin \frac{2\pi d \sin \theta}{\lambda} \sum_n (\tilde{A}_n \cdot d\tilde{E}_n) \right]^2. \end{aligned} \quad (11)$$

To interpret this, we notice that $\sum_n (\tilde{A}_n \cdot d\tilde{E}_n) = d\Sigma(\tilde{A}_n \cdot \tilde{E}_n) = d(\Sigma E_n^2) = dI_0$, i.e. the contribution from dS to the illumination at A .

We now introduce an approximation. It is a well-known result of geometrical optics that if a source of uniform brightness is observed through an optical system of any kind, some part or the whole of the aperture of the system is seen by an eye focused on it to be illuminated with a brightness equal to that of the source. If the whole of the aperture is so illuminated the condition is known as a 'complete flash'; otherwise, as a 'partial flash'.

The assumption is now made that the region S' of the aperture which is filled with light, as determined by geometrical optics, does not vary greatly for points in the region of A and B . In the majority of cases likely to be met with in practice this is true. Then the illumination I_0 on the object plane is equal to $S'B$, where B is the surface brightness of the source. Also, $dI_0 = BdS$. Hence, the terms $\sum_n (\tilde{A}_n \cdot d\tilde{E}_n)$ in (11) can be replaced by BdS , the associated integrals being taken over S' .

To interpret the terms $\sum_n \tilde{A}_n \times d\tilde{E}_n$, the 'principle of stationary phase' may be evoked. This states that the value $\tilde{E}_n$ of the resultant at A of the n th wave-front is contributed chiefly by those elements of area dS for which $d\tilde{E}_n$ has a stationary value of phase, and that $\tilde{E}_n$ agrees in phase with those $d\tilde{E}_n$.

It may be seen, therefore, that a given element dS will, for some wave-fronts, lie in the areas of stationary phase, and that $\tilde{E}_n$ will have the same phase as $d\tilde{E}_n$ for these wave-fronts. Accordingly, $\tilde{A}_n \times d\tilde{E}_n = \tilde{E}_n \times d\tilde{E}_n = 0$ for such wave-fronts. For the remainder of the wave-fronts $\tilde{E}_n$ and $d\tilde{E}_n$ will have mutual phase-angles varying rapidly with the distance of dS from the corresponding region of stationary phase. Hence the values of $\tilde{A}_n \cdot \tilde{E}_n$ will be in the long run just as much positive as negative. It follows that $\sum_n \tilde{A}_n \times d\tilde{E}_n$ will be small as compared with $\sum_n \tilde{A}_n \cdot d\tilde{E}_n$.

Equation (11) may therefore be written as

$$M^2 = \left[\int \cos \frac{2\pi d \sin \theta}{\lambda} B dS \right]^2 + \left[\int \sin \frac{2\pi d \sin \theta}{\lambda} B dS \right]^2 \\ = \left[B \int_{S'} \cos \frac{2\pi d \sin \theta}{\lambda} dS \right]^2 + \left[B \int_{S'} \sin \frac{2\pi d \sin \theta}{\lambda} dS \right]^2. \quad (12)$$

Hence, from (5),

$$V = \frac{M}{I_0} = \frac{B \left\{ \left[\int_{S'} \cos \frac{2\pi d \sin \theta}{\lambda} dS \right]^2 + \left[\int_{S'} \sin \frac{2\pi d \sin \theta}{\lambda} dS \right]^2 \right\}^{\frac{1}{2}}}{I_0} \\ = \frac{1}{S'} \left\{ \left[\int_{S'} \cos \frac{2\pi d \sin \theta}{\lambda} dS \right]^2 + \left[\int_{S'} \sin \frac{2\pi d \sin \theta}{\lambda} dS \right]^2 \right\}^{\frac{1}{2}}. \quad (13)$$

The right-hand side of (13) is immediately recognizable as the ratio of the amplitudes at the central maximum and at a point distant d therefrom in the diffraction pattern of a perfect lens, the aperture of which has the same shape and angular subtense as the area S' of the condenser aperture which is filled with light.

This obviously reduces to the two cases considered by Van Cittert when the condenser is perfect and sharply focused, or when no condenser is present at all, and equation (13) then gives the same results as Van Cittert found. As was found by Zernike, if the source is so big that the condenser aperture appears to be filled with light, the coherence in the object plane is independent of the aberrations or condition of focusing of the condenser. Hence, the state of correction of a microscope condenser is of importance only if it is desired to use a small light source.

If, then, the two fields obtained in the arrangement described in the previous section are not quite in coincidence but are so placed that there is a distance d between corresponding points in the two fields, and interference effects take place which would be expected, if the coincidence were perfect, to have a visibility of unity as defined above, the actual visibility will be given by

$$V = \frac{J_1(2\pi d n \sin \alpha / \lambda)}{\pi d n \sin \alpha / \lambda}, \quad (14)$$

where α is the semi-angle of the illuminating cone, which is supposed circular, and n is the refractive index in the object space. The small part of the illuminating cone stopped out by the central silvered spot has been ignored, and the assumption made that the condenser is 'completely flashed', in (14). This is the well-known expression for the amplitude in the diffraction pattern of a circular aperture, and it follows that to obtain good visibility the error of coincidence d must be materially less than the resolving limit which corresponds to the numerical aperture of the illuminating cone. This means that the correct relative alinement of the two interferometer plates is very critical, so the lower plate must be provided with a well-designed set of levelling screws.

PRACTICAL ARRANGEMENT

The first experimental model was mounted on an old Zeiss microscope which had a rotating stage with centring adjustments. The necessary modifications to the stand were very small, consisting only of the addition of a small steel plate with three

levelling screws which was mounted in the well of the stage. The lower interferometer plate was mounted in a hole in the centre of this steel plate. The necessary translational motion was then obtained by means of the centring adjustments of the stage.

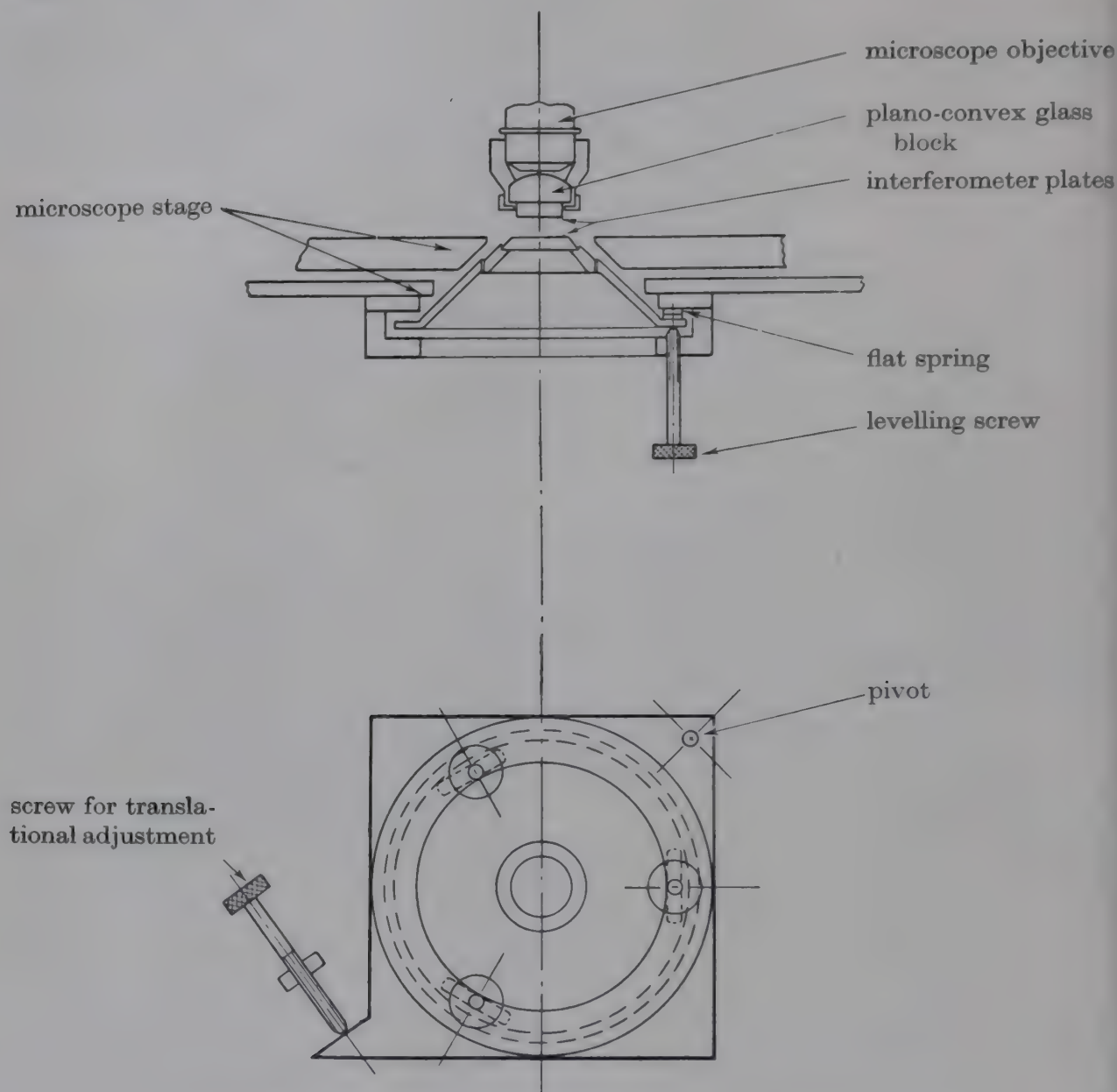


FIGURE 5. Mechanical arrangement of interferometer microscope.

This arrangement, although it gave fair results, proved somewhat inconvenient in use, and the apparatus was later mounted on a modern Beck stand which had a built-in mechanical stage. All the necessary movements had now to be provided in the additional apparatus. The rotational adjustment was provided by mounting the objective, together with the upper interferometer plate and plano-convex glass block, on a 'loose ring' which could be clamped once the necessary adjustment was obtained. The lower plate, which was approximately $\frac{3}{4}$ in. in diameter and 0.080 in. thick, was mounted at the top of a thin-walled truncated steel cone which extended down below the stage and which was supported at its base by three levelling screws working against beryllium-copper leaf springs. The levelling screws were in turn

mounted on a steel plate about 3 in. square, which was pivoted at one corner and held in contact with the underside of the stage by springs. This could be rotated slightly about the pivot by a screw bearing against the corner opposite the pivot and working against springs. This provided the translation motion for the bottom interferometer plate.

The condenser was mounted on a somewhat lengthened tube and was accommodated inside the steel cone. An ordinary Abbé condenser with iris diaphragm was used in all the experiments described below.

This arrangement is shown diagrammatically in plan and elevation in figure 5, the plan view being drawn looking upwards at the underside of the stage.

OPERATIONAL PROCEDURE

Assuming that all the adjustments are wrong initially, the apparatus may be set up as follows. The lower interferometer plate is raised or lowered by the levelling screws until it is about 0.5 mm. below the top surface of the stage. A drop of immersion oil is placed on it and the slide placed on the stage. The oil must be sufficient in quantity to make contact with the bottom of the slide. A second drop of oil is placed on top of the slide and the microscope tube racked down till the oil makes contact with the lower surface of the upper plate. The illumination is centred and cut down to a rather narrow cone by means of the iris. The object is now roughly focused; it will usually be possible to ascertain the point of approximate focus if the iris opening is small.

In general, two non-coincident images of the light source will be seen in the field of view. They are brought to coincidence, as nearly as can be estimated by the eye, by adjustment of the levelling screws. The iris is now opened out, the eyepiece removed and the illuminating aperture of the objective inspected. It will, in general, be seen to be traversed by a system of approximately straight equidistant fringes. These are caused by interference between the wave-fronts originating from corresponding points in the two images and are similar in origin to the Fizeau fringes observed between a pair of flat surfaces inclined to each other at a small angle. Hence, these fringes form a sensitive test for coincidence or otherwise of the two images. By means of the levelling screws, the fringes are made to spread out until the aperture appears illuminated with uniform intensity. Upon replacing the eyepiece the field will, in general, appear to be crossed by a series of equidistant fringes, the spacing depending on the angle between the wedge axes of the two plates. The object will appear superimposed on the fringes, exhibiting a type of contrast varying with its position on the fringe system.

Usually the best observing conditions are obtained when the path-difference between the two interfering beams is uniform over the field, i.e. when the fringe spacing is infinite. This is obtained by rotating the objective together with the upper plate and repeating the above adjustments. Two or three repetitions of this procedure will suffice to bring the wedge axes sufficiently nearly parallel.

Operation of the screw giving translational motion (the 'phase screw') when several fringes are in the field causes the fringe system as a whole to advance across

the field in a direction at right angles to the fringes. If white light be used, the fringes exhibit the colour sequence of Newton's rings and, as the path-difference may be made of either sign, there is a central bright fringe which is white and indicates that the path-difference is zero along that fringe. If the fringe spacing is infinite the whole field exhibits colour changes simultaneously.

PHASE AND AMPLITUDE CONTRAST

We represent the amplitudes in the two interfering beams by E_1, E_2 . If the phase difference be θ , then the resultant E is given by

$$E^2 = E_1^2 + E_2^2 + 2E_1 E_2 \cos \theta. \quad (15)$$

The intensity I will be proportional to E^2 .

If an object which has no absorption but introduces a small phase-difference $d\theta$ is present, the change of intensity will be

$$dI = d(E^2) = \frac{\partial(E^2)}{\partial\theta} d\theta = -2E_1 E_2 \sin \theta d\theta. \quad (16)$$

We define the 'contrast' C as $C = \frac{dI}{I}$. (17)

Then $C = -\frac{2E_1 E_2 \sin \theta d\theta}{E_1^2 + E_2^2 + 2E_1 E_2 \cos \theta} = -\tan \frac{1}{2}\theta d\theta$, (18)

when $E_1 = E_2$. It follows that the contrast is zero for objects of small optical thickness if $\theta = 0$, i.e. if the intensity is a maximum, and is infinite if $\theta = \pi$, which corresponds to dark field conditions.

To consider the case when the two interfering beams are not equal, let $E_2 = kE_1$. Then the contrast is given by

$$C = -\frac{2k \sin \theta}{1 + k^2 + 2k \cos \theta} d\theta, \quad (19)$$

which is a maximum when $\cos \theta = -\frac{2k}{1+k^2}$. Substituting this back into (19), the maximum value of the contrast is given by

$$C = -\frac{2k}{1-k^2} d\theta. \quad (20)$$

The values of the coefficient of $d\theta$ for a few values of k are given in the table below:

k	$-\frac{2k}{1-k^2}$
0.9	- 9.5
0.95	- 19.5
0.98	- 49.5
0.99	- 99.6
1.00	∞

It is evident that if it is desired to detect very small differences of optical thickness it is important to pay particular attention to equality of the two interfering beams. As the conditions for high sensitivity correspond nearly to dark-field conditions, it is also essential that scattered light be kept to a minimum.

We now consider an object which exhibits no difference of optical path but which has a slight absorption, reducing the intensity E_2^2 in one of the interfering beams to $E_2^2 - d(E_2^2)$. Then

$$d(E^2) = \frac{\partial(E^2)}{\partial E_2} \frac{dE_2}{d(E_2^2)} d(E_2^2) = \frac{E_2 + E_1 \cos \theta}{E_2} d(E_2^2)$$

and

$$C = \frac{E_2 + E_1 \cos \theta}{E_2(E_1^2 + E_2^2 + 2E_1 E_2 \cos \theta)} d(E_2^2) = \frac{d(E_2^2)}{2E_2^2}, \quad (21)$$

when $E_1 = E_2$. This is one-half of the contrast which would be obtained by examination of such an object under conditions of normal microscopy and is independent of the phase angle between the two fields when the two beams are equal.

EFFECT OF THICKNESS OF OBJECT

If the object is of thickness t and refractive index n , immersed in a medium of refractive index n_0 , it introduces a change of optical path equal to $t(n - n_0)$. In addition, however, it causes a relative shift, parallel to the axis, between the two interfering fields of amount $t(n - n_0)/n$, and if this shift is too great the two fields will become incoherent. The order of magnitude of the allowable shift can be obtained from equation (13). It will be noted that in the derivation of (13) no restriction was placed on the direction of the line AB , and this can be taken as parallel to the axis of the instrument. This being so, the right-hand side of (13) represents the ratio of the amplitudes at two points distant d apart on the axis, and will be sensibly unity if d does not exceed the physical depth of focus corresponding to the illuminated aperture of the condenser. Assuming a circular aperture of semi-angle α , the depth of focus is equal approximately to $\pm \lambda/2n_0\alpha^2$. The complete range over which the two amplitudes are sensibly equal is thus $\lambda/n_0\alpha^2$.

Accordingly, interference effects of good visibility will be obtained if the thickness t does not exceed a value given by

$$t(n - n_0)/n = \lambda/n_0\alpha^2,$$

or

$$t = \frac{\lambda n}{\alpha^2 n_0(n - n_0)}. \quad (22)$$

This corresponds to a path difference of $\lambda n/n_0\alpha^2$, or to a number of fringes equal to $n/n_0\alpha^2$. For $n = 1.5$, $n_0 = 1.33$ and $\alpha = 30^\circ$ this corresponds approximately to 4.5 fringes. Thus, the thickness of an object may be sufficient to give several fringes before a perceptible loss of contrast ensues.

EXPERIMENTAL RESULTS

The first object to be examined in the interferometer microscope was a suspension of a fine grade of the abrasive 'Aloxite' in immersion oil. The grains have a refractive index differing appreciably from that of the oil and are very irregular in shape. Figure 6, plate 4, is a photograph (magn. $\times 300$) taken under dark-field conditions. It will be seen that some of the grains have an optical thickness sufficiently great for several fringes to appear within their boundaries. These form a contour map of the thickness of the grain; a very clear impression of the form of the surfaces can be obtained by operating the phase screw and observing the march of the fringes across

the image. The nature of some detail of which the interpretation is obscure can often be elucidated by watching the behaviour of the moving fringes in its neighbourhood. This facility is one of the most valuable features of the interferometer method and is one of its chief advantages over the conventional form of phase contrast.

As the path difference between the two interfering beams can be varied continuously, the white fringe of zero path-difference can be moved from the background to any given portion of the object and, by counting the number of fringes, the optical thickness can be estimated. In a later section a method of measuring fractional parts of a fringe of optical thickness will be described.

Figure 7, plate 4, shows a piece of thin collodion membrane of the type used as the specimen support in electron microscopy. This was cast on a water surface, then gathered up on to a microscope slide without any special precautions, resulting in a puckered and folded layer as the photograph indicates. The superimposed folds of membrane illustrate clearly the manner in which contrast depends on optical thickness, and the complete absence of any 'halo effect' is very evident.

Figure 8, plate 4, shows the edge of a single layer of such a film. The thickness was later measured by Tolansky's technique of multiple-beam interferometry and proved to be approximately 180 Å. No special precautions had been taken in this instrument to make the two interfering beams equal, yet it can be seen that a film of this thickness gave a contrast which was readily capable of being photographed. The degree of contrast seen visually may be estimated from the fact that the edge of the film was parallel to the short edge of a 3 × 1 in. slide and, on 'sweeping' the slide from one end under a power of 500, the edge was noted the first time it was crossed.

Small flakes of mica are beautiful objects when viewed in this manner. The edge of a flake torn from the parent sheet has a terraced formation which shows up as a series of bands of varying intensity or colour. As the microscope is working with full aperture and unrestricted angle of illuminating cone, the resolution is good and minute cracks may sometimes be seen in the surface.

The appearance of such terraced areas varies from the very complex structure of figure 9, plate 5, to the surprising regularity shown in figure 10, plate 5. It will be seen that although the wave-length of the 'saw-teeth' of figure 10 varies slightly across the field, the minute details, such as the slight overhang of the white tips of the teeth, are reproduced faithfully from tooth to tooth. The series of teeth ran for a distance equal to about three times the width of the field, or 0.6 mm., the pitch decreasing by about one-half from one end of the series to the other.

A phenomenon was seen on several mica flakes which at first was very puzzling. This was the appearance of terraces which seemed to have a curved cross-section as shown diagrammatically in figure 11*a*, plate 5, as evidenced by a variation of colour or intensity across each terrace in a direction at right angles to its length. This seemed scarcely compatible with the structure of the mica, but the mystery was resolved when these variations were observed to behave in a somewhat peculiar manner when the focus was altered. It turned out that the true structure was as shown in figure 11*b*, plate 5, with terraces of complementary nature and equal depth of steps on the upper and lower surfaces of the mica. The thickness of the mica is therefore constant, the variations of intensity observed on the upper surface being



FIGURE 6. 'Aloxite' grains in immersion oil (magn. $\times 300$).



FIGURE 7. Collodion film (magn. $\times 300$).



FIGURE 8. Collodion film 180 Å thick (magn. $\times 300$).



FIGURE 9. Mica flake (magn. $\times 300$).



FIGURE 10. Mica flake (magn. $\times 300$).

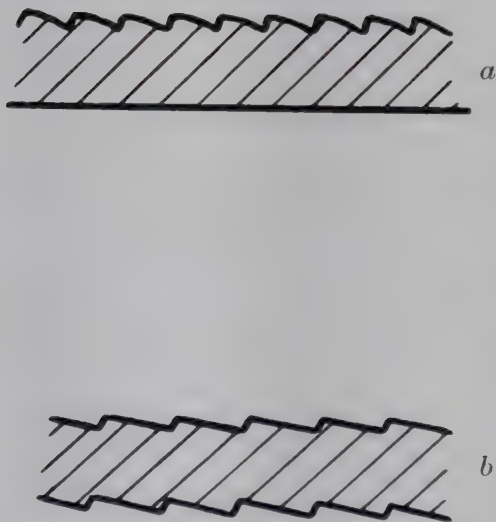


FIGURE 11. Apparent and real cross-section of mica flake.

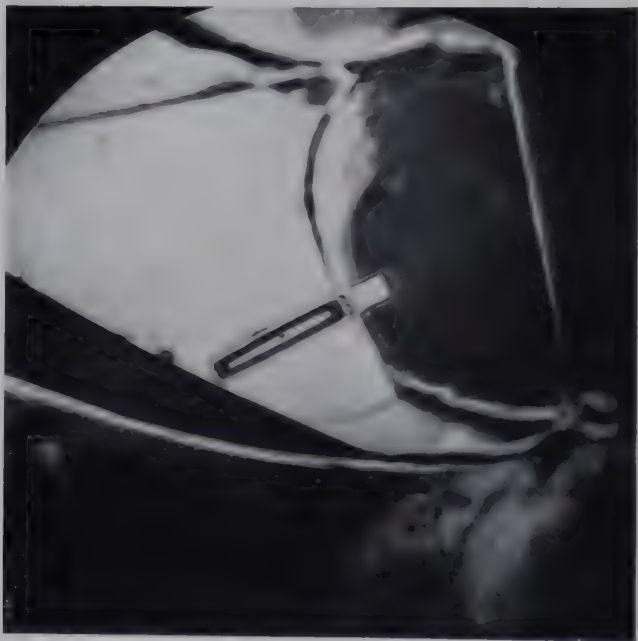


FIGURE 12. Unknown object on mica flake (magn. $\times 450$).

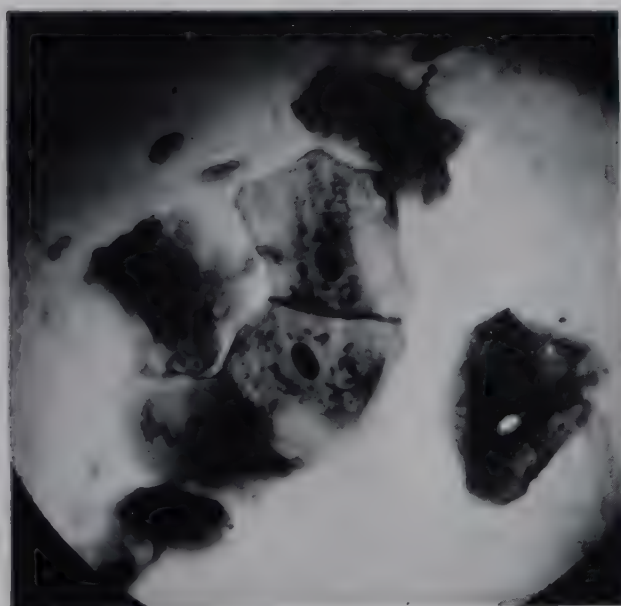


FIGURE 13. Epithelial cells (magn. $\times 300$).

FIGURE 14. Epithelial cells (magn. $\times 300$).

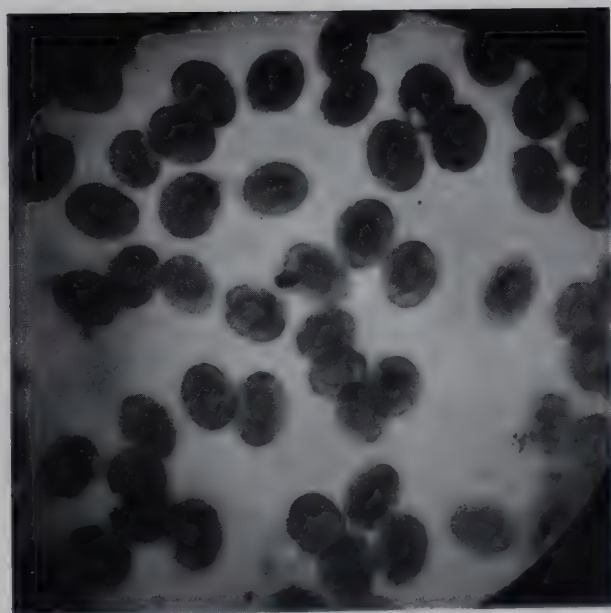


FIGURE 15. Frog's blood cells in immersion oil (magn. $\times 300$).

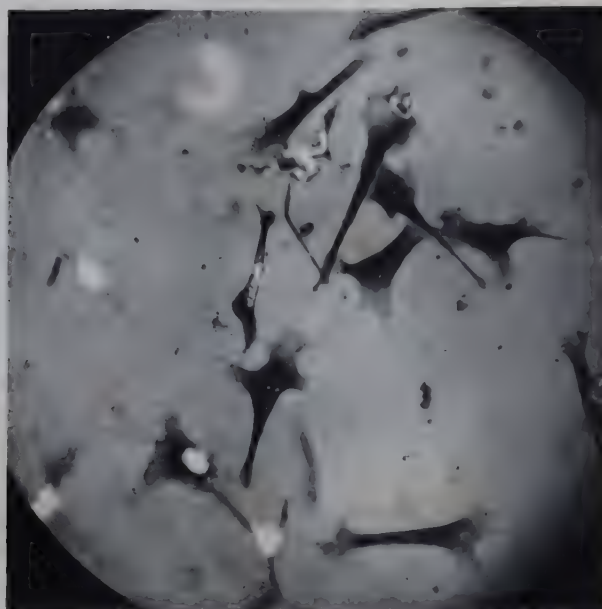


FIGURE 16. Snail amoebocytes grown in tissue culture (magn. $\times 300$).

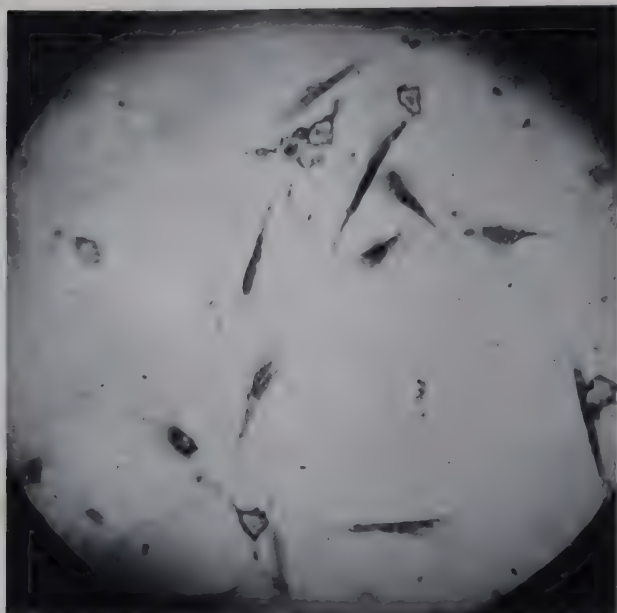


FIGURE 17. Snail amoebocytes grown in tissue culture (magn. $\times 300$).

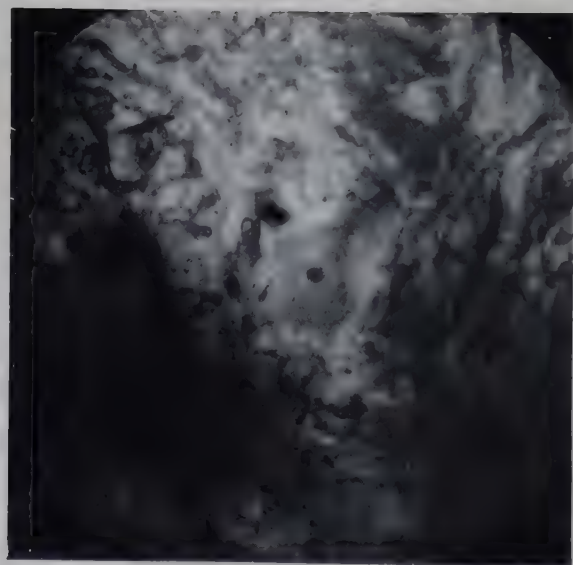


FIGURE 18. Monkey's spinal cord without interference contrast (magn. $\times 300$).

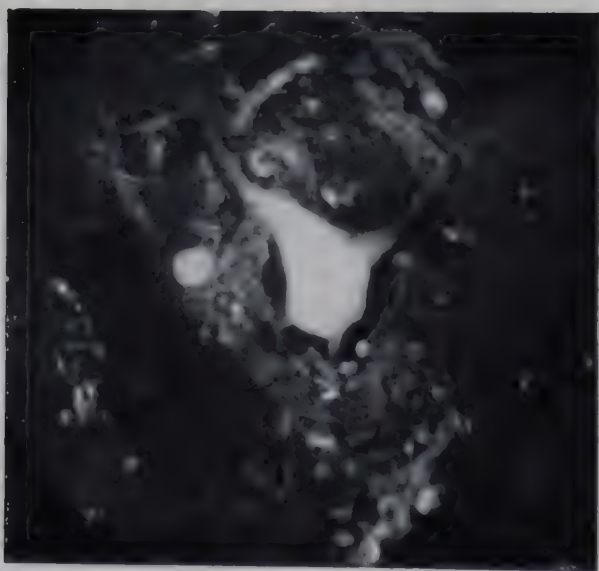


FIGURE 19. Monkey's spinal cord with interference contrast (magn. $\times 300$).

caused by the out-of-focus light diffracted from the lower terrace steps. It was possible to focus at will on either the upper or lower terraces and to confirm that the steps on the two surfaces were exactly alined. The existence of this phenomenon has been inferred by Tolansky.

Figure 12, plate 5, shows another mica flake, but the interest of this photograph lies in the cylindrical object which is resting on the surface of the mica. The irregularity near one end is apparent only, being due to the terrace on the underlying mica being seen through the object which is quite transparent and is of very nearly uniform section. It resembles a short length of capillary tubing with a diameter of 7μ . The object is immersed in water, and about two-thirds of the length of the 'bore' is seen to be occupied by what appears to be an air-bubble. By using the central white fringe as a 'marker' it was possible to show, by varying the path-difference, that the object was in fact of smaller optical thickness in the space occupied by the bubble than at other places along its axis. The maximum optical thickness was about two fringes of green light, which assuming a circular cross-section, corresponds to a refractive index of 1.49. This object has not been identified.

The interferometer microscope has been tried out on a number of biological objects. Figures 13 and 14, plate 6, show a small group of epithelial cells photographed with different settings of the path-difference, corresponding to bright-field and dark-field conditions. Figure 15, plate 6, shows a preparation of frog's blood cells in immersion oil, which are rather difficult objects by phase contrast. Figures 16 and 17, plate 7, show snail amoebocytes grown in tissue culture. These illustrate the flexibility of the interference method, for figure 16 is taken with a phase setting such as to give maximum contrast of the extremely thin outer sheets of cytoplasm, causing the nuclei to be practically blacked out, whereas figure 17 shows a different phase setting to bring out some details of the nuclei.

Finally, figures 18 and 19, plate 7, show a stained section of monkey's spinal cord. These were taken to ascertain how serious a limitation is the requirement that the object should be small as compared with the cross-section of the comparison beam. Figure 18 was taken with the levelling screws not in correct adjustment and hence no interference is taking place. In figure 19, which shows the same field, the adjustment is correct, and it is obvious from the appearance of the large cell in the centre that interference is occurring. This could be checked by operation of the phase screw, when the customary changes of contrast occurred. The object is several millimetres across, and so the whole cross-section of the comparison beam is occupied by both phase and amplitude detail of a density comparable with that shown in the photographs. Hence, the necessity for smallness of the object is not such a drawback as at first sight might appear. All the photographs shown here were taken with a 4 mm. objective, the numerical aperture of the system being 0.75, with a condenser aperture in the region of 0.5.

MEASUREMENT OF OPTICAL THICKNESS

As the phase screw is operated the brightness of any portion of the image varies according to the expression $a + \cos p$, p being the path-difference in radians and a being, ideally, unity. The optical thickness of any portion can be measured by

ascertaining the difference between the epochs of the cosine term for the portion in question and for the background. The nature of a cosine curve prevents this being done accurately simply by observing the brightness variation, but an imperfection in the apparatus suggested a way of doing this with precision.

If the lower interferometer plate be raised by the levelling screws to bring the silvered spot on its lower surface into focus simultaneously with the object any imperfections in its surface appear superimposed on the object. A number of fine scratches about 5μ wide were observed, and these were of nearly rectangular cross-section with flat bottoms. Within such a scratch the path-difference differs by a fraction of a micron from that in the general field, and so the brightness of the scratch varies with a different epoch from the brightness of the field. For a given setting of the phase screw the scratch can be made to disappear; this setting is extremely critical, the brightness of scratch and field varying in opposite directions for a small motion of the phase screw.

Accordingly, to measure the optical thickness of a detail, a scratch is selected of which the length is parallel to the direction in which the plate is moved by the phase screw. The object is moved till the scratch passes through the detail; then the phase adjustment is varied till the scratch disappears against the detail, or the detail disappears against the scratch if the former is very small. The position of the silvered spot is read off on an eyepiece graticule, using some easily recognized irregularity of the scratch, and the phase varied again to make the scratch disappear against the background of a portion of the field which is clear. Its position is noted and the difference between the two graticule readings taken. Finally, the distance of motion between two successive disappearances of the scratch against the background is measured.

The above is done in monochromatic light. The ratio of the two readings gives the optical thickness as a fraction of a complete wave-length of path-difference, with an ambiguity of any whole number of wave-lengths. The ambiguity can now be removed by using white light and noting the amount of traverse of the phase plate required to bring the central white fringe from the background to the detail being investigated.

Using this method a determination of the refractive index of the cytoplasm of epithelial cells from the tongue was attempted. The chief factor tending to reduce the accuracy was the extreme difficulty of measuring the dimensional thickness, which was done by focusing on the lowest and highest observable detail in the cell and measuring the difference in height by the fine focusing knob of the microscope. This gave a wide scatter for the measurement of the thickness even of a single cell, for the total thickness is only a few microns. However, as the cells were immersed in water and only the comparatively small difference in index between cell and water had to be measured, this did not affect the result as much as might be expected. The final result was 1.358 ± 0.010 , the limits representing the two most divergent results obtained on measurement of six cells.

The scratch employed was not perfectly regular, but a precision of about $\frac{1}{40}$ of a fringe could be obtained in the setting for disappearance. No doubt better results could be obtained by using a diamond scratch of controlled form. An alternative

would be to use a line scraped through the silver deposit, with an opaque backing to prevent the passage of the direct light from the condenser. This would give an area of one-quarter the maximum brightness of the field which did not vary with the phase adjustment. This could also be achieved by depositing the specimen on a slide covered with a network of fine opaque lines. Both these methods should theoretically give a smaller precision of setting than that obtainable with the method first described, but as they do not depend upon uniformity of depth of a scratch they might in practice give better results.

CONCLUSIONS

An interference method for the microscopical examination of transparent objects has been described which is free from some of the limitations of conventional phase contrast microscopy and which can give additional information. It differs from most other methods of interferometer microscopy in that it can be used on an ordinary microscope stand without major modifications and without restriction on the aperture of illumination. There is a restriction on the size of the object being examined, but an example has shown that this is not so serious as might be imagined.

A disadvantage of the method is the large loss of light occasioned by the many reflexions at half-silvered surfaces, but even so, using a B.T.H. 250 W mercury-vapour box lamp, it is possible to photograph on 35 mm. Microfile film in about $\frac{1}{5}$ sec. when using integrated light and about 2 sec. when using an Ilford 'mercury-green' filter. This would give ample speed for time-lapse cinemicrophotography on to 16 mm. film.

The chief value of the new method would seem to lie in the ease with which conditions can be changed to give maximum visibility of any detail in which the observer is interested and the possibility of making accurate quantitative measurements of optical thickness and, in some cases, of refractive index. In addition, extra sensitivity may be obtained by using white light and the 'sensitive-tint' technique. Experience alone will decide the value of the method to the practical microscopist.

In conclusion, the author wishes to acknowledge the assistance given during this development by Dr R. Barer in the form of much helpful discussion and criticism, and to thank him for supplying most of the biological objects of which photographs are reproduced above. In addition, the author wishes to thank Dr T. E. Allibone, F.R.S., for permission to publish this paper.

REFERENCES

- Bates, W. J. 1947 *Proc. Phys. Soc.* **59**, 940.
- Dyson, J. 1949 *Proc. Phys. Soc. B*, **62**, 565.
- Frederikse, A. M. 1933 *Acta brev. neerl. Physiol.* **3**, 8.
- Merton, Sir Thomas 1947a *Proc. Roy. Soc. A*, **189**, 309.
- Merton, Sir Thomas 1947b *Proc. Roy. Soc. A*, **191**, 1.
- Van Cittert, P. H. 1934 *Physica*, **1**, 201.
- Zernike, F. 1938 *Physica*, **5**, 785.
- Zernike, F. 1948 *Proc. Phys. Soc.* **61**, 158.

Paramagnetic resonance in chromic sulphate alums at room temperature

BY D. M. S. BAGGULEY AND J. H. E. GRIFFITHS

Clarendon Laboratory, University of Oxford

(Communicated by B. Bleaney, F.R.S.—Received 31 May 1950—

Revised 24 July 1950)

Five of the chromic sulphate alums have been investigated by the method of paramagnetic resonance at room temperature. Except for the potassium alum, a crystalline electric field of trigonal symmetry gives an adequate explanation of the positions and intensities of the lines in the absorption spectrum. The observed widths of the absorption lines correspond closely with those calculated from magnetic dipole interaction.

1. INTRODUCTION

Measurements of specific heats and relaxation phenomena at low temperatures have shown that the ground state of most paramagnetic salts includes a system of closely spaced energy levels. Further evidence has been derived from the interpretation of paramagnetic susceptibilities. The measurement of the absorption spectrum of such solids at centimetre wave-lengths, however, allows a direct investigation of these levels. In this paper are described measurements on various chromic sulphate alums, all of which have the same type of absorption spectrum, and an interpretation is given in terms of the energies and the lowest states of the Cr^{+++} ion under the influence of the internal crystalline electric field, and the applied magnetic field. Since the absorption spectrum gives a direct measure of the energy difference between two states at a given magnetic field, a detailed comparison between theory and experiment should be possible. Unfortunately, the absorption lines have an appreciable width and often are not fully resolved, so that in some cases the comparison cannot be made as accurately as could be wished. On the whole the agreement is good, particularly in the Rb, Cs and (NH_3CH_3) alums where the initial splitting is larger than in the potassium salt and the absorption lines more clearly resolved. An accurate value for the Stark splitting of the energy levels in zero magnetic field can also be derived from these results; this is discussed in §4.

The broadening of the absorption lines has been investigated theoretically by Van Vleck (1948). The width is assumed to be caused by simple dipolar coupling between the paramagnetic ions, and the effect of the crystalline electric field is neglected. The widths measured experimentally are in good agreement with this theory.

From the known wave functions of the Cr^{+++} ion under the influence of the assumed electric and magnetic fields, it is possible to calculate the transition probabilities of the various absorption lines. The calculation has been performed by Professor M. H. L. Pryce and Mr G. A. Wheatley (private communication) for the two particular cases in which the magnetic field is directed along the normals to (100) and (111) planes in the crystal. The measured value of the Stark splitting

enables the position of the absorption lines to be calculated so that it is possible to construct a theoretical spectrum if the half-widths and line shapes are known. Assuming a Gaussian line shape and the theoretical half-widths good agreement has been obtained with the theoretical curves.

2. GENERAL EXPERIMENTAL PROCEDURE

The experimental method and arrangement is described in some detail in this paper, since only a brief account has been published previously (Bagguley & Griffiths, 1947, 1948; Bagguley, Bleaney, Griffiths, Penrose & Plumptre, 1948).

The method adopted is to place the paramagnetic substance in a tuned circuit, situated between the poles of a magnet, and measure the change in Q (circuit magnification) as the applied magnetic field is varied, the incident frequency meanwhile remaining constant. At centimetre wave-lengths the tuned circuit is either a coaxial line or a cavity resonator. The wave in a high Q cavity may be regarded as traversing the enclosed space many times before its final decay, and so a very large effective path length is thereby obtained. By choosing an H mode resonator the paramagnetic salt may conveniently be placed where the electric field is small.

When estimating the magnitude of the absorption to be detected it is convenient to do so in terms of the Q of the cavity resonator, which is of the order 5000 at centimetre wave-lengths. Using the relations given by Kittel & Luttinger (1948), and assuming the loaded Q to be 1000 when the salt is in the resonator at a wave-length of 3 cm., a change in Q of about 10 % due to resonance absorption would be expected in our experiments. This is, of course, an ideal case for lines having only dipolar broadening. When the line width is greater the peak absorption intensity decreases, maintaining the total area under the curve constant. The sensitivity can be improved only to a limited extent by increasing the amount of paramagnetic salt in the resonator, since additional dielectric losses are introduced which reduce the initial Q of the cavity. It may be shown that if $Q_0^{(s)}$ is the circuit magnification of the resonator with salt, and Q_0 is the corresponding quantity for the resonator without the salt, then the optimum quantity of salt is that for which $(Q_0^{(s)}/Q_0)$ is $\frac{1}{3}$.

Q is defined to be the ratio of energy stored in the resonator to the energy lost per radian and may be measured by detuning the resonator to the half-power points. However, if Q_0 is the circuit magnification when there is no paramagnetic absorption, and Q is the value of this quantity at a magnetic field H , then the energy loss at this field strength due to resonance absorption is proportional to $(Q_0/Q - 1)$. This can be measured most conveniently by the power transmitted through the cavity when it is tuned to resonance. For this purpose a silicon-tungsten crystal detector is coupled to the resonator, and the rectified current from the crystal detector (square-law characteristic) is proportional to Q^2 . If δ_0 and δ are the galvanometer deflexions corresponding to Q_0 and Q , the paramagnetic absorption is proportional to $(\sqrt{(\delta_0/\delta)} - 1)$, and the absorption curve is obtained by plotting this function against magnetic field H .

The use of high Q resonators demands good frequency stability, since small changes in frequency detune the cavity. Sufficient stability was obtained by supplying all the potentials for the oscillator valve, a reflexion klystron, from a stabilized power

supply. The voltages were constant to 1 part in 10^4 , and the galvanometer deflexion (100 cm.) remained constant to within 1 mm. or at worst 1 cm. for the duration of a set of measurements (20 to 30 min.).

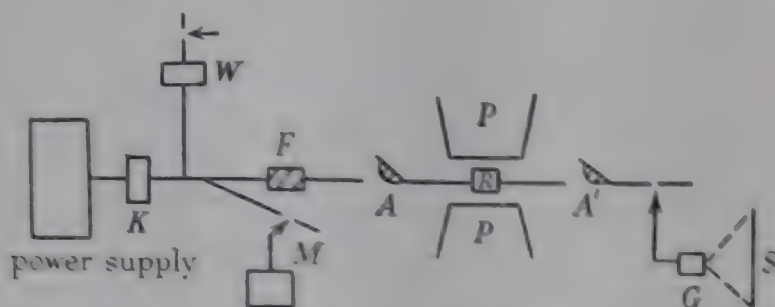


FIGURE 1. Block diagram of apparatus.

A block diagram of the apparatus is shown in figure 1. The power from the reflexion klystron K was fed through rectangular wave guide (at the longer wave-lengths a coaxial line was used) of length 2 m., into a resonator R , placed between the pole faces P of an electromagnet. A wavemeter W and a directive feed terminated in an untuned silicon-tungsten crystal rectifier for use as a power monitor M , were attached to the guide near the oscillator. F was a fixed attenuator, whilst A , A' were variable. The fixed attenuator (about 7 db.) was sufficient to prevent reflected waves from the resonator affecting the valve, whilst A and A' were used to obtain a convenient power level for the measurements. It was verified that with a total length of 2 m. of waveguide stray field from the magnet had no effect on the oscillator valve. The sensitivity of the galvanometer G which measured the crystal current was 90 cm./ μ A with the scale S at a distance of 3 m. The steady magnetic field was arranged to be perpendicular to the oscillating magnetic field in the resonator, being produced between flat pole faces about 8 cm. in diameter having a separation of 2.5 cm. The magnet was supplied with current from the 100 V d.c. mains, the voltage of which was usually steady to within 1%. A relation between exciting current and magnetic field was established using a Grassot fluxmeter and search coils standardized at the N.P.L. Field strengths less than 6000 G were estimated to be measured with an accuracy better than 1%, but above this the calibration was probably not accurate to more than 2%. Measurements were always taken with the magnet working under the same conditions, and it was found that the calibration was accurately reproducible over considerable periods of time.

Crystals of K, Rb, Cs and (NH_4CH_3) chromic sulphate alums were grown by slow evaporation of a solution placed in a thermostat which was maintained at 25°C to within 0.1°C. The edges of the crystals were at least 1 cm., so that a section cut parallel to a (110) plane had an area of about 1 sq.cm. The sections were from 1 to 3 mm. thick. This plane contains the normals to the (100), (110) and (111) type planes, and so by placing the section in the bottom of the resonator, these directions could be set at known angles to the steady magnetic field by simple rotation of the crystal section. With such large crystals it was often impossible to obtain sections free from cracks. However, it was verified by using sections from many different crystals that these defects did not affect the experimental results.

3. DETAILS OF THE APPARATUS

At wave-length of about 3 cm.

In this wave-length region the oscillator was of the type CV 318 or CV 87, supplying between 50 and 100 mW of power. The valve was coupled by coaxial line to rectangular waveguide having cross-section $2\frac{1}{2} \times 1$ cm.; this was excited in the H_{01} mode by a probe. The wavemeter W enabled the wave-length to be measured with an accuracy of 1 part in 10^3 , and the monitor M allowed power fluctuations greater than 2% to be detected. The resonator consisted of an accurately bored cylindrical brass tube, 2.20 cm. in diameter constant to within 0.001 cm. throughout its length. The cavity was excited in the H_{11} mode and was coupled to the line through two irises, situated $\frac{1}{4}$ of a wave length from one end, each being 4 mm. in diameter in a wall of 1.5 mm. thickness. One end of the resonator was closed by a head which was attached to the spindle of a micrometer screw. This head had a conventional filter section to reduce the effect of uncertain contacts between the head and the walls. A plunger with a similar choke system closed the other end, and on this was placed the paramagnetic crystal section. A graduated scale on the plunger enabled the crystal to be rotated through known angles and set in given directions with respect to the magnetic field.

The resonator was silver plated, and the measured Q value without paramagnetic salt was 10,000 compared with a theoretical value of 17,000 at 3.5 cm. wave-length.

With this type of tuning mechanism the resonator could easily and accurately be retuned. This was done before each measurement of the galvanometer deflexion.

At wave-lengths between 4 and 10 cm.

Since the cut-off wave-length of the resonator described above was 3.75 cm., a larger resonator had to be used at the longer wave-lengths; in order to retain the gap between the pole faces of the magnet at 2.5 cm. a rectangular resonator was necessary. It was also more convenient to convey the high-frequency power by concentric line rather than by waveguide. Pyrotex cable having an attenuation of 3 db. m. at 10 cm. wave-length was used. The oscillators used were a NS 2 in the 4 to 5 cm. range and a CV 308 which could be tuned from 5 to 10 cm. The cross-sectional dimensions of the resonator were 6.3×2.5 cm. It was tuned by a rectangular plunger with filter sections, driven by a micrometer screw through a ball-race. This arrangement was not entirely satisfactory since there was a small amount of backlash, and also the filter section was not so successful in removing the effects of uncertain contacts. For these reasons the resonator was not retuned before each reading, but it was found that this did not affect the results. Coupling to the resonator was by means of loops formed directly on the coaxial line. The bottom of the resonator was closed by a piece of brass which was held in place by screws through the side of the resonator. This base-plate had a circular hole 2.2 cm. diameter in the centre through which was inserted the plunger on which the specimen was placed. Except at 6.7 cm., where a circular mode necessitated the use of a distrene plunger, the plunger was of brass and had a simple choke system. A graduated scale enabled the crystal

to be rotated through known angles. The whole resonator was silver plated and without paramagnetic salt the measured Q was 3000, compared with a theoretical Q of 10,000 at 6 cm. wave-length.

At wave-lengths of about 1 cm.

Power from a reflexion klystron type VX 302 was transmitted along 2 m. of rectangular waveguide into a cylindrical resonator 2.20 cm. diameter excited in the H_{01} mode. Two small irises each 2.5 mm. in diameter in a wall of 0.5 mm. thickness enabled power to be fed through the cavity. One end of the resonator was closed by a non-contact plunger attached to the spindle of a micrometer screw. This head was backed by lossy material to remove any modes which might be excited behind the plunger. A similar non-contact plunger, on which the paramagnetic crystal was placed, screwed into the bottom of the resonator. The crystal could be rotated through known angles and set in given directions with respect to the steady magnetic field. The unloaded Q value for an H_{01} resonator of these dimensions at 1.25 cm. is 12,000. The measured value without paramagnetic salt was 10,000.

The resonator was retuned before each reading of the galvanometer deflexion, but fluctuations in power transmitted through the resonator due to frequency and amplitude instability of the valve were more serious in these measurements than those made at the longer wave-lengths. The drift in some experiments corresponded to 10 % of the peak resonant absorption intensity.

4. THE CRYSTALLOGRAPHY OF THE ALUMS

A detailed analysis of the crystal structure of the alums has been given by Beevers & Lipson (1934). Although all the alums are based on space group Pa^3 , (T^6h), slight differences in atomic arrangement enable three classes to be distinguished, denoted by Lipson (1935) as the α , β and γ types. Practically all the data given by the above authors refer to alums having Al^{+++} as the trivalent ion; they do, however, include both $KAl(SO_4)_2 \cdot 12H_2O$ and $KCr(SO_4)_2 \cdot 12H_2O$ as having the same structure. The change in unit-cell dimensions is only 1 part in 300, so that it seems reasonable to assume that the slight difference in radius between the two trivalent ions does not affect the classification. This is confirmed by the external symmetry of the crystals. The alums that have been examined in the experiments described in this paper are:

α alums	β alums
$KCr(SO_4)_2 \cdot 12H_2O$	$(NH_3CH_3)Cr(SO_4)_2 \cdot 12H_2O$
$NH_4Cr(SO_4)_2 \cdot 12H_2O$	$CsCr(SO_4)_2 \cdot 12H_2O$
$RbCr(SO_4)_2 \cdot 12H_2O$	

The structure differences are apparently due to the variation in size of the monovalent ion; in the β structure the very large Cs^+ and $(NH_3CH_3)^+$ ions are in contact with six oxygen atoms of SO_4^- groups in addition to six water molecules. There is a difference also in the octahedra of water molecules surrounding the trivalent ion. For alums of the α type, the X-ray evidence reveals a distortion along a trigonal axis of unit cell, together with a rotation of 12° about this direction. The β alums, however, appear to have perfectly regular groups of water molecules, the axes of the octahedron

being directed along those of unit cell. It follows that for α -type alums unit cell will contain four inequivalent chromic ions, each having the distortion of the surrounding octahedron of water molecules along one of the trigonal axes of the crystal. These ions are situated at (000) , $(\frac{1}{2}\frac{1}{2}0)$, $(\frac{1}{2}0\frac{1}{2})$ and $(0\frac{1}{2}\frac{1}{2})$. In the case of the β alums, however, the X-ray evidence suggests that the four Cr^{+++} ions are equivalent.

Theory

The energy levels of the Cr^{+++} ion in potassium chrome alum have been investigated by Van Vleck (1939) and Broer (1942). The crystalline field is assumed to consist of two parts, one with cubic symmetry and a smaller component having trigonal symmetry. The axis of the trigonal field is directed along one of the trigonal axes of unit cell.

The ground state for Cr^{+++} is ^4F and the effect of the cubic field is partially to lift the orbital degeneracy, splitting the state into a basic singlet and two levels with threefold multiplicity. Addition of the trigonal component, together with the spin orbit coupling, separates the triplets and resolves the spin quadruplet into two levels having quantum numbers $\pm \frac{3}{2}$, $\pm \frac{1}{2}$ along the trigonal axis. The energy difference between the two spin levels, in zero magnetic field, is of the order 0.1 cm.^{-1} , whereas the separation between the two lowest orbital levels is about $10,000 \text{ cm.}^{-1}$. At room temperature, therefore, only the singlet orbital level is populated, and transitions may be induced between the four spin levels resolved by an external magnetic field.

Using this model, Broer has shown that the energy of a spin level is given by one of the roots of the equation

$$W^4 + 2aW^3 + W^2(a^2 - 10\mu_0^2H^2) - 2aW\mu_0^2H^2(7 - 6c) + 9\mu_0^4H^4 - a\mu_0^2H^2(4 - 3c) = 0,$$

in which a is the initial Stark splitting of the spin levels, μ_0 is the Bohr magneton, $c = \cos^2\theta$ being the angle between the applied magnetic field H and the trigonal axis.

When $c = 1$ the equation can be solved rigorously, giving

$$W_1 = -a + 3\mu_0H,$$

$$W_2 = +\mu_0H,$$

$$W_3 = -\mu_0H,$$

$$W_4 = -a - 3\mu_0H.$$

For other values of c the equation can conveniently be considered as a quadratic in H^2 , when graphical solutions may be obtained for all the cases required.

The four inequivalent Cr^{+++} ions in unit cell will have different values of c for an arbitrary orientation of the external magnetic field. However, if the magnetic field is directed along one of the space diagonals of unit cell, one Cr^{+++} ion will have $c = 1$, whilst three will have $c = \frac{1}{9}$. In figure 2 the energies of the spin levels for these two cases have been plotted. In the absorption experiments, transitions between two energy levels are detected, so that in comparing theory with experiment it is useful to plot values of energy difference between two levels $W_n - W_m$ as a function of magnetic field. Figures 3, 4 and 5 give for potassium chrome alum, therefore, the position of the absorption lines as a function of magnetic field strength for the three

directions normal to (111), (110) and (100) planes. In order to calculate these curves, the initial Stark splitting was assumed to be 0.12 cm.^{-1} ; they may be applied to any chromic alum if the scales are increased in the ratio of the Stark splittings.

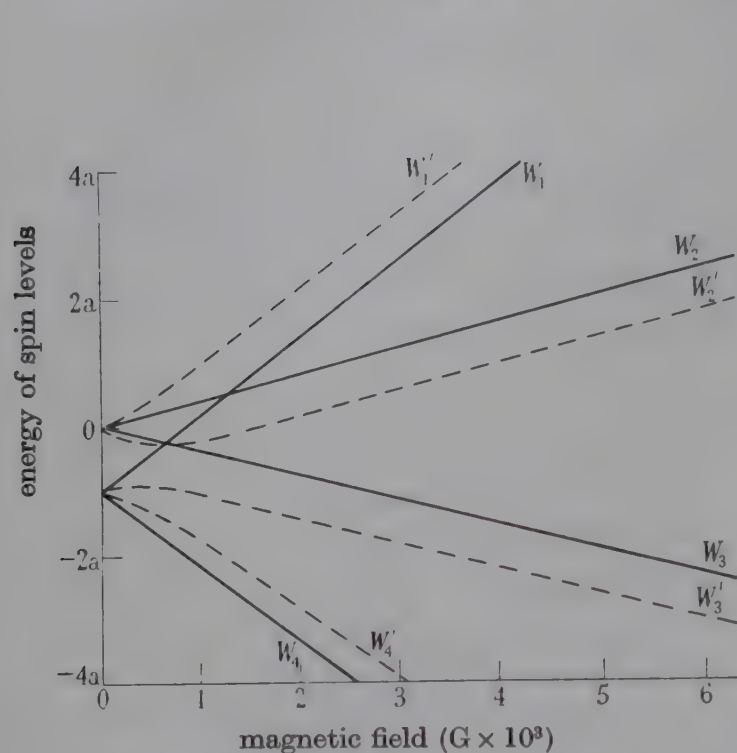


FIGURE 2. Energy of spin levels for magnetic field in (111) direction.

— $c = 1$; --- $c = \frac{1}{9}$

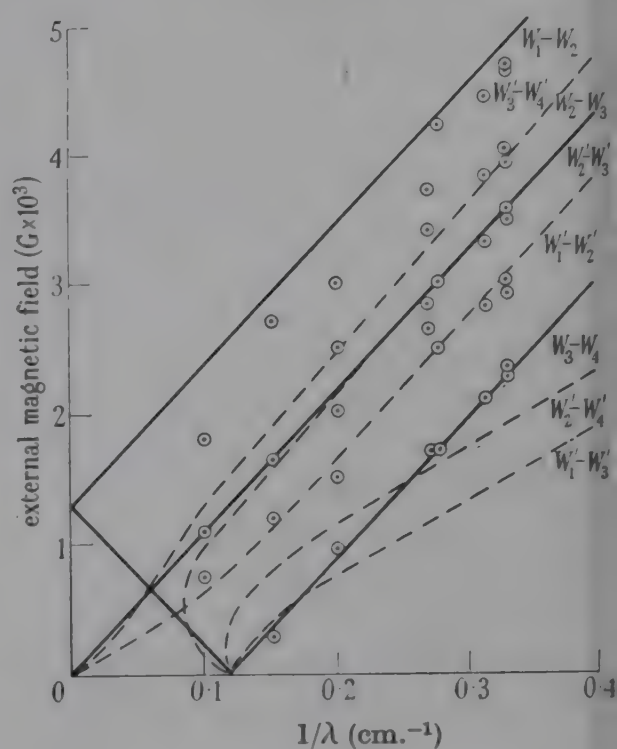


FIGURE 3. Energy differences for $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$, (111) direction.

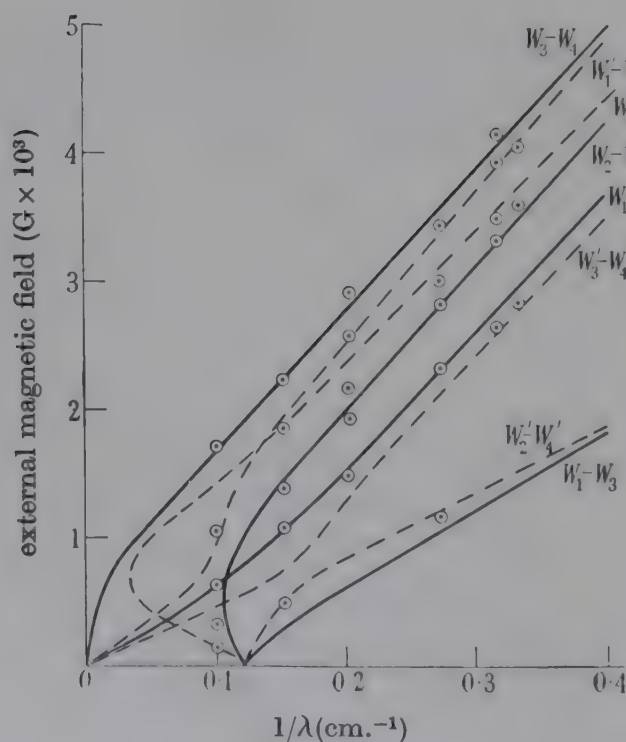


FIGURE 4. Energy differences for $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ (110) direction. — $c = 0$; --- $c = \frac{2}{3}$.

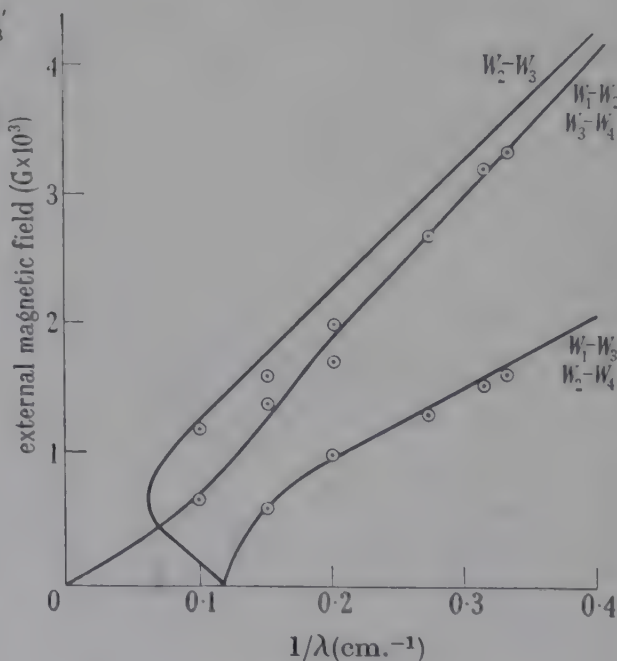


FIGURE 5. Energy differences for $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ (100) direction. $c = \frac{1}{3}$.

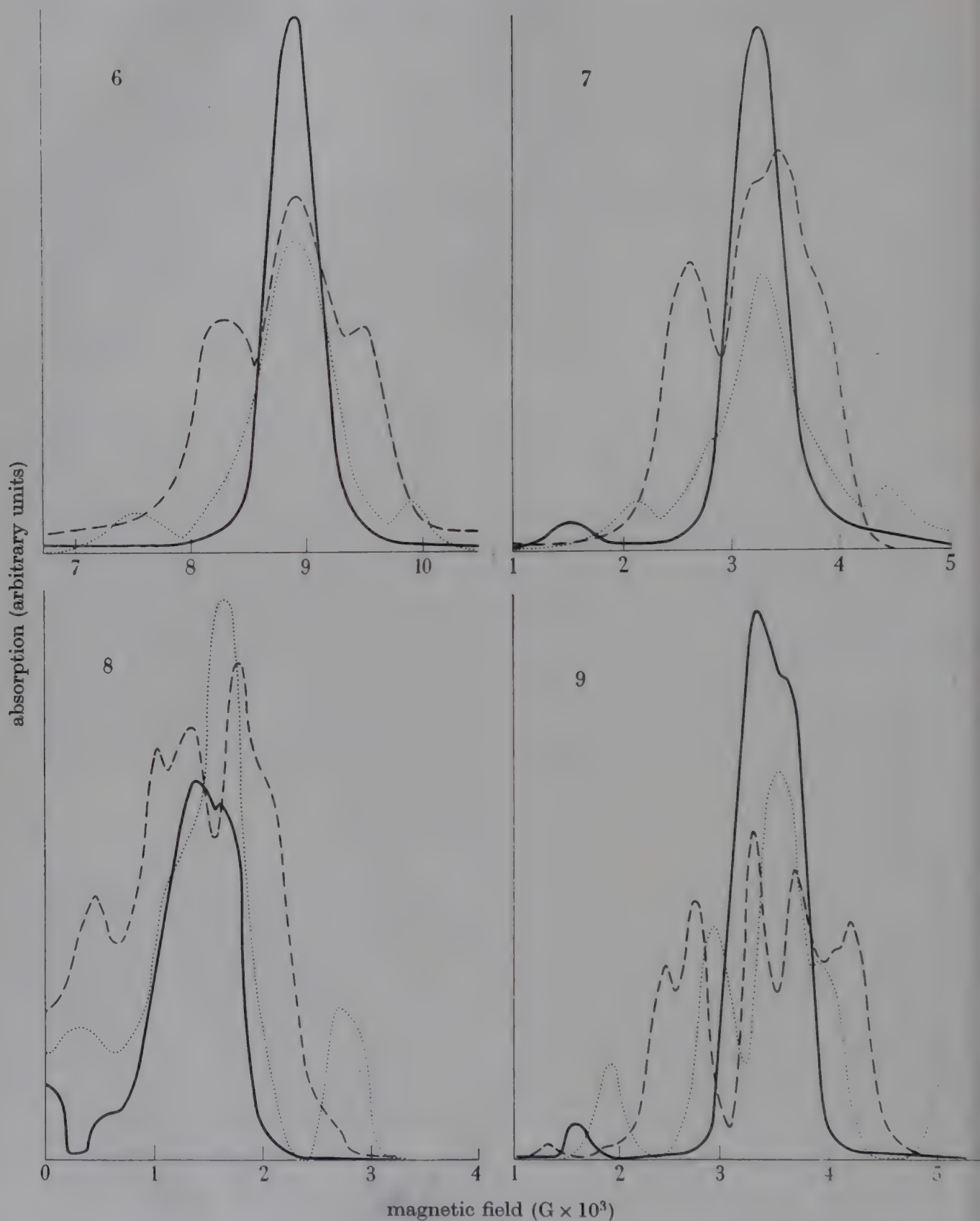
5. EXPERIMENTAL RESULTS

Measurements have been made on various chromic alums at room temperature and typical absorption spectra for $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at wave-lengths of 1.23, 3.18 and 6.56 cm. are shown in figures 6, 7 and 8. In each figure three curves are given corresponding to the normals to (111), (110) and (100) planes being in the direction of the steady magnetic field. The absorption spectrum of $\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at a wave-length of 3.02 cm. is given in figure 9. The other alums have similar spectra, the resolution of the different absorption lines improving as the initial Stark splitting of the spin levels increases. In the figures the applied magnetic field is plotted as abscissa and the ordinate is $(\sqrt{(\delta_0/\delta)} - 1)$ (δ being the galvanometer deflexion) which is proportional to the absorption. No attempt was made to obtain an absolute value for the absorption, nor to make a comparison between different orientations or different wave-lengths, and therefore the ordinate scales cannot be compared directly.

The curves for the same orientation but different wave-lengths have much the same form, but are shifted to higher field strengths at the shorter wave-lengths. At the longer wave-lengths the form of the curves does change somewhat, since some of the absorption lines occur near zero magnetic field, so that at 10 cm. the absorption in large fields is less than that in zero field. Here the ratio of volume of crystal to that of the resonator is very small and hence the absorption only just detectable, the results are then correspondingly less accurate. The Stark splittings deduced from the absorption spectra of the different alums are given in table 2. Measurements of the ammonium and potassium alums have been published by Whitmer, Weidner, Hsiang & Weiss (1948) and by Halliday & Wheatley (1948). For these salts our results are in agreement with the latter workers, whilst the discrepancy with the other measurements may be due to the particular experimental arrangement adopted by these authors.

Discussion of the different alums $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$

The absorption spectrum of potassium chrome alum was first investigated, and was followed throughout the whole range of wave-length 1.23 to 10 cm. In figures 3, 4 and 5 the positions of the observed absorption lines are compared with theory for the three orientations (111), (110) and (100), the points being the observed positions of the absorption lines. The initial Stark splitting was taken to be 0.12 cm.^{-1} , and this appears to give reasonable agreement between experiment and theory. For the magnetic field directed along the (111) direction there is an asymmetry in the extreme absorption peaks, the line occurring in high fields being displaced towards the central peak. This may perhaps be due to a non-trigonal component of the crystalline electric field, or to a displacement of this trigonal field from the (111) direction in the crystal. On the latter assumption an angular displacement of 17° would be required, but there is no crystallographic evidence in support of this hypothesis. In many cases the resolution is insufficient for the individual lines to be distinguished; thus for the (100) direction of the three transitions corresponding to a change in spin quantum number of unity, two are coincident, and the third lies too close to be resolved in the experimental curves.



FIGURES 6 to 9. Absorption spectra of: 6, $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 1.23 cm.; 7, $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 3.18 cm.; 8, $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 6.56 cm.; 9, $\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 3.02 cm. — (100); --- (110); (111).

$\text{NH}_4\text{Cr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$

This alum has been investigated at wave-lengths of 3.57, 3.02, 1.78 and 1.23 cm. Although the measured Stark splitting of 0.135 cm.^{-1} is greater than for the potassium alum, it is still insufficient for the absorption lines to be completely resolved. In table 1 the position of the absorption lines for the magnetic field in the (111) directions are compared with theory. For the other directions the agreement was to the same order of accuracy.

TABLE 1. ANALYSIS OF THE ABSORPTION SPECTRA

wave-length (cm.)	orientation		position of lines (Gauss)
(NH ₄)Cr(SO ₄) ₂ · 12H ₂ O {Stark splitting 0.135 cm. ⁻¹ }			
3.57	(111)	experiment	1460, 2380, 3000, 3360, 4380
		theory	1540, 2440, 3000, 3360, 4460
3.02	(111)	experiment	2100, 3020, 3540, 3860, 4940
		theory	2100, 3080, 3540, 3940, 4940
1.25	(111)	experiment	7370, 8200, 8630, 9200, 10,000
		theory	7120, 8050, 8570, 9000, 10,000
1.79	(111)	experiment	4550, 5550, 6070, 6500, 7450
		theory	4550, 5470, 6000, 6430, 7430
CsCr(SO ₄) ₂ · 12H ₂ O {Stark splitting 0.145 cm. ⁻¹ }			
3.02	(111)	experiment	2020, 2980, 3550, 3900, 5060
		theory	2040, 2980, 3550, 3920, 5060
3.57	(111)	experiment	1460, 2380, 3050, 3400, 4500
		theory	1520, 2440, 3020, 3400, 4520
RbCr(SO ₄) ₂ · 12H ₂ O {Stark splitting 0.165 cm. ⁻¹ }			
3.12	(111)	experiment	1630, 2710, 3420, 4170, 5170
		theory	1650, 2710, 3420, 3750, 5190
(NH ₃ CH ₃)Cr(SO ₄) ₂ · 12H ₂ O {Stark splitting 0.165 cm. ⁻¹ }			
3.04	(111)	experiment	1770, 2730, 3510, 3975, 5300
		theory	1770, 2760, 3530, 3960, 5300

The other alums

The Rb, Cs and (NH_3CH_3) alums have been investigated at 3 cm. wave-length only. The comparison between experiment and theory becomes more exacting as the Stark splitting increases, since, as the line width remains practically constant, the absorption lines are more clearly resolved. In all cases agreement between the theory and experiment is good, the assumption of a trigonal field acting on the chromic ions providing an adequate explanation of the observed spectra, to the accuracy of these experiments. The position of the absorption lines for the magnetic field in the (111) direction are compared with theory in table 1. For other orientations of the crystal similar agreement was obtained.

To explain the positions of some absorption lines occurring in low magnetic fields, it is necessary to assume that transitions corresponding to a change in spin quantum number of two or three occur. The observed intensities are in agreement with this assumption (see § 7).

From the structure parameters and sizes of unit cell given by Beevers & Lipson (1934) it is possible to calculate the position of the six water molecules surrounding

the trivalent ion. Unfortunately, practically all the data given by these authors refer to Al^{+++} alums, and experiments on dilute chromic alums have shown that substitution of Al^{+++} for Cr^{+++} produces sufficient distortion of the lattice to change the Stark splitting of the spin quadruplet by more than 20 %. The change in position of the water molecules necessary for such an effect may, however, be too small to be detected by X-ray measurements. Table 2 gives some information about the Al^{+++} alums corresponding to those which we have investigated. The third column gives the distances between adjacent water molecules belonging to the group surrounding the trivalent ion, which are a measure of the distortion of the octahedra; whilst column 4 gives the separation between the ion and its nearest neighbours. In the last column the observed Stark splitting is listed for the corresponding chromic alum; the X-ray measurements do not appear to be simply related to this quantity.

TABLE 2

substance	unit-cell dimensions a (Å)	$\text{H}_2\text{O}-\text{H}_2\text{O}$ (Å)	$\text{Al}-\text{H}_2\text{O}$ (Å)	Stark splitting in Cr alum (cm^{-1})
$(\text{NH}_3\text{CH}_3)\text{Al}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	12.479	2.65, 2.65	1.87	0.165 ± 0.005
$\text{CsAl}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	12.333	2.79, 2.79	1.98	0.145 ± 0.005
$\text{RbAl}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	12.220	2.79, 2.83	1.99	0.165 ± 0.005
$(\text{NH}_4)\text{Al}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	12.215	2.79, 2.83	1.99	0.135 ± 0.005
$\text{KAl}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	12.133	2.77, 2.81	1.97	0.12 ± 0.005

We remarked earlier that for β type Al^{+++} alums the X-ray evidence suggests that the octahedron of water molecules surrounding the Al^{+++} ion is perfectly regular, with axes along those of unit cell. However, the similarity of the absorption spectra for all the alums that we have examined indicates that there is present in the β type Cr^{+++} alums a trigonal electric field of sufficient magnitude to produce a splitting of the order 0.15 cm^{-1} in the basic spin quadruplet. The estimates of Van Vleck (1939) show that the Coulomb field from distant atoms is unlikely to cause a splitting of more than 0.05 cm^{-1} , so that the trigonal field must arise principally from a distortion of the water octahedron. It appears, therefore, that a distortion of the water octahedron does exist in β type Cr^{+++} alums and is of the same order of magnitude as that found in the α alums.

It should also be remarked that the measured splittings at room temperature are not in agreement with estimates from relaxation and specific heat measurements below 1°K . The low-temperature measurements give 0.16 to 0.19 cm^{-1} for potassium chromic alum (Bleaney 1950; Gorter 1947) and 0.34 cm^{-1} for the ammonium chromic alum (Starr 1941). Good agreement between measurements at different temperatures is not, however, to be expected, since for most Cr^{+++} alums the Stark splitting varies considerably with change of temperature; a detailed account of this variation is given in the following paper by B. Bleaney.

6. LINE SHAPE AND HALF-WIDTHS

When the Stark splitting of the basic spin quadruplet is greater than 0.13 cm^{-1} , the absorption lines corresponding to transitions between levels $W_1 - W_2$ and $W_3 - W_4$, for those ions with $c = 1$, become completely resolved. For the crystal sections used

in these experiments, the maximum intensity of the absorption for these lines corresponded to a change in Q of the resonant cavity by about 5 %. It was possible, therefore, to delineate the absorption curve fairly accurately. Other absorption lines, corresponding to higher order transitions, are usually completely resolved, but the intensity is too small for accurate measurement to be made. The experimental points obtained with a crystal of $\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at a wave-length of 3.04 cm. have been compared with the Gaussian curve. The wings of the observed lines do appear more intense than would be expected for a pure Gaussian curve, but the measured effect is here small, and agreement is to within the accuracy of the experiments.

The absorption measurements also enable the line width at half-maximum intensity to be measured. Since the basic orbital level for the Cr^{+++} ion is a singlet, and the separation from the lowest excited triplets state is about $10,000 \text{ cm.}^{-1}$, spin-lattice effects may be expected to be small, and the line broadening should be caused by dipole-dipole interaction. This is confirmed by measurements at low temperatures (Bleaney 1950) which show that the change in line width between room temperature and 20° K is not more than 10 G.

Van Vleck (1948) has shown that if the line breadth is small compared with the separation of the energy levels, and effects due to the crystalline field are neglected, the mean square local field due to dipolar interaction is given by

$$H_i^2 = \frac{1}{3}(g\mu_0)^2 s(s+1) \sum_j (1 - 3 \cos^2 \theta_{ij}) r_{ij}^{-6},$$

in which g is the Landé g factor, μ_0 is the Bohr magneton, s is the resultant spin of each ion, θ_{ij} is the angle between the direction of the steady magnetic field and the line joining the two dipoles, and r_{ij} is the distance between two dipoles. The contributions to H_i^2 are multiplied by a factor $\frac{3}{4}$ if the ions are equivalent.

The half-width of the absorption lines (assumed Gaussian) have been calculated for two particular orientations of the crystals in the magnetic field. It has been assumed that the factor $\frac{3}{4}$ should be included for all pairs of ions, having the same 'c' values. When the steady magnetic field is along a (100) direction, $c = \frac{1}{3}$ for all the Cr^{+++} ions, and the half-width is given by

$$H_{\frac{1}{2}} = 28.5 \times 10^{-20} a^{-3} \text{ G},$$

in which a is the length of an edge of unit cell in cm. For the magnetic field directed along the normal to a (111) plane, $c = 1$ for one ion and $c = \frac{1}{9}$ for three ions in each unit cell. The corresponding half-widths are

$$c = 1: \quad H_{\frac{1}{2}} = 25.2 \times 10^{-20} a^{-3} \text{ G},$$

$$c = \frac{1}{9}: \quad H_{\frac{1}{2}} = 33.7 \times 10^{-20} a^{-3} \text{ G},$$

and a is again the length of an edge of unit cell.

In table 3 the half-widths of the absorption lines derived from these relations are given, together with those obtained from the absorption spectra. In applying the theoretical values of $H_{\frac{1}{2}}$ to the different alums it has been assumed that the dimensions of unit cell correspond to those for the Al^{+++} alum, except in the case of $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$. The accuracy of the experimental measurements is, in some cases, not very high, owing to incomplete resolution of the individual lines, and in the

table mean values are given, obtained from the complete set of experimental curves. The agreement between theory and experiment is good, the most serious discrepancy occurring for $\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$, but in this case the lines are not very clearly resolved. Simple dipolar interaction, therefore, appears to provide an adequate mechanism for broadening the absorption lines in the chromic alums.

TABLE 3. THE HALF-WIDTHS (IN G) OF ABSORPTION LINES
IN THE Cr^{+++} ALUMS

substance	$c = \frac{1}{3}$	$c = 1$	$c = \frac{1}{9}$	
$\text{KCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	158	140	187	theory
	200	200	—	experiment
$\text{NH}_4\text{Cr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	156	138	185	theory
	—	170	190	experiment
$\text{RbCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	156	138	185	theory
	150	140	200	experiment
$\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	152	134	180	theory
	180	140	180	experiment
$(\text{NH}_3\text{CH}_3)\text{Cr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$	147	130	173	theory

The dielectric losses in methylamine alum were very great, so that only a very thin crystal section could be used. The absorption intensity was then too small for the line breadth to be estimated.

7. TRANSITION PROBABILITIES

From the known wave functions of the Cr^{+++} ion under the influence of the assumed electric and magnetic fields, it is possible to calculate the transition probabilities of the various absorption lines. This calculation has been performed by Professor M. H. L. Pryce and Mr G. A. Wheatley (private communication), for the two particular cases in which the magnetic field is directed along the normals to (100) and (111) planes in the crystal.

It has been remarked previously that the experimental curves are consistent with a Gaussian line shape, and also allow determination of the half-width of an absorption line. The measured value of Stark splitting enables the position of the lines to be derived, and so using the calculated transition probabilities it is possible to construct the absorption spectrum. This has been performed for $\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ in the (111) and (100) directions at 3.02 and 3.57 cm. wave-length, and for $\text{RbCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ in the (100) direction at 3.47 cm. wave-length.

For the magnetic field along a normal to a (111) plane in $\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$, the agreement between theory and experiment is good. In figure 10 the experimental points and theoretical curve are plotted together. The Stark splitting was assumed to be 0.145 cm.^{-1} , and the line shape was such that $H_{\frac{1}{2}} = 140 \text{ G}$ for the ions having $c = 1$, and $H_{\frac{1}{2}} = 180 \text{ G}$ for those with $c = \frac{1}{9}$. Good agreement could not be obtained by assuming that all the lines had the same width. In constructing the absorption spectrum it was found that the maximum intensities of four lines were practically unaffected by the wings of others. These can be calculated from the transition probabilities, if allowance is made for the different widths of the lines, and a comparison between theory and experiment is given in table 4. There is serious discrepancy in only one instance, that of $\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$, for 3.57 cm. wave-length

and $H = 3040$ G. Here the measured intensity was small, and the valve rather unsteady, so that the error is most probably in the experimental results.

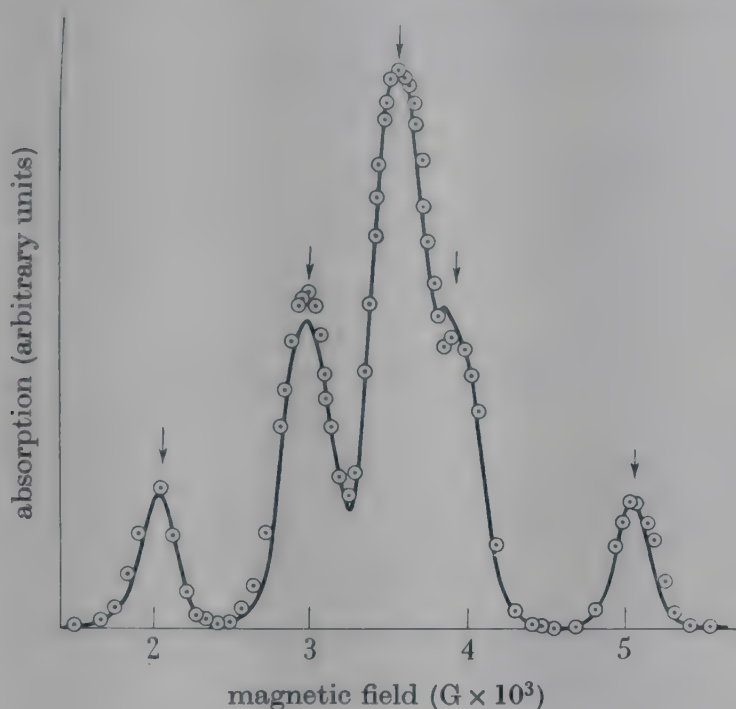


FIGURE 10. Theoretical absorption spectrum for $\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 3.02 cm. with experimental points.

TABLE 4. A COMPARISON BETWEEN THE THEORETICAL AND OBSERVED MAXIMUM INTENSITIES OF RESOLVED ABSORPTION LINES

Unit intensity has been chosen arbitrarily. Steady magnetic field along normal to (111) plane.

$\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 3.02 cm.

position of line (G)	2040	2980	3560	5060
theoretical intensity	1	2.3	4.2	1
experimental intensity	1	2.4	4.2	1

$\text{CsCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 3.57 cm.

position of line (G)	1500	2400	3040	4500
theoretical intensity	1	2.3	4.0	1
experimental intensity	1	2.2	3.5	1

$\text{RbCr}(\text{SO}_4)_2 \cdot 12\text{H}_2\text{O}$ at 3.12 cm.

position of line (G)	1650	2700	3420	5240
theoretical intensity	1	2.1	3.7	1
experimental intensity	1	2.0	3.6	1

A similar analysis has been performed for the absorption spectrum observed when the applied magnetic field is directed along the normal to a (100) type plane. In this case higher order transitions of the type $W_1 - W_3$ and $W_2 - W_4$ have a reasonably high probability at the longer wave-lengths, and the corresponding absorption lines have been detected. In table 5 the calculated and observed intensities are given for those lines in the spectrum for which the peak is practically unaffected by the wings of neighbours. The agreement is again within the experimental error.

TABLE 5. A COMPARISON BETWEEN THE THEORETICAL AND OBSERVED MAXIMUM INTENSITIES OF RESOLVED ABSORPTION LINES

Unit intensity has been chosen arbitrarily. Steady magnetic field along normal to (100) plane.

CsCr(SO ₄) ₂ · 12H ₂ O at 3·57 cm.			
position of line (G)	700	1400	2800
theoretical intensity	0·25	1	5·8
experimental intensity	0·2	1	6·2
RbCr(SO ₄) ₂ · 12H ₂ O at 3·12 cm.			
position of line (G)	780	1600	3170
theoretical intensity	0·26	1	5·9
experimental intensity	0·2	1	5·9
RbCr(SO ₄) ₂ · 12H ₂ O at 3·42 cm.			
position of line (G)	—	1430	2750
theoretical intensity	—	1	5·0
experimental intensity	—	1	5·3

8. CONCLUSION

We have described measurements of the absorption spectra of various chromic alums at room temperature and centimetre wave-lengths. The results indicate that for both the α and β chromic alums, the octahedron of water molecules surrounding the trivalent ion is sufficiently distorted to produce a splitting of the basic spin levels varying from 0·12 to 0·165 cm.⁻¹. In the case of Rb, (NH₄), Cs and (NH₃CH₃) chromic alums the assumption of a trigonal electric field having its axis along the normal to a (111) plane of the crystal provides good agreement between theory and experiment. Unfortunately, such an assumption does not appear to give a complete explanation of the absorption spectrum for the potassium salt. The experimental values for the line widths are consistent with the theory of simple dipolar coupling given by Van Vleck (1948); whilst the agreement between the calculated and observed absorption intensities affords additional proof of the adequacy of the trigonal electric field, and for it being directed along a space diagonal of unit cell. It is also a test of the particular value of Stark splitting chosen, and it seems that the values given are correct to within the accuracy of the present experimental arrangement, i.e. 2 or 3 %.

Our thanks are due to Professor Lord Cherwell for extending to us the facilities of the Laboratory; to many of our colleagues for very helpful discussions, in particular Professor M. H. L. Pryce and Dr B. Bleaney; and to the Board of Admiralty for apparatus and other assistance in carrying out the work. One of us (D. M. S. B.) was assisted during the course of these experiments by a grant from D.S.I.R.

REFERENCES

- Bagguley, D. M. S., Bleaney, B., Griffiths, J. H. E., Penrose, R. P. & Plumpton, B. I. 1948 *Proc. Phys. Soc.* **61**, 542.
 Bagguley, D. M. S. & Griffiths, J. H. E. 1947 *Nature*, **160**, 532.
 Bagguley, D. M. S. & Griffiths, J. H. E. 1948 *Nature*, **162**, 538.
 Beever, C. A. & Lipson, H. 1934 *Proc. Roy. Soc. A*, **146**, 570.
 Bleaney, B. 1949 *Phys. Rev.* **75**, 1962.

- Bleaney, B. 1950 *Proc. Roy. Soc. A*, **204**, 203.
 Bleaney, B. & Penrose, R. P. 1948 *Proc. Phys. Soc.* **60**, 395.
 Broer, L. J. F. 1942 *Physica*, **9**, 547.
 Gorter, C. J. 1947 *Paramagnetic relaxation*. New York: Elsevier Publishing Co. Inc.
 Halliday, D. & Wheatley, J. 1948 *Bull. Amer. Phys. Soc.* **23**, no. 3, 13.
 Kittel, C. & Luttinger, J. M. 1948 *Phys. Rev.* **73**, 162.
 Lipson, H. 1935 *Proc. Roy. Soc. A*, **151**, 347.
 Starr, C. 1941 *Phys. Rev.* **60**, 241.
 Van Vleck, J. H. 1939 *J. Chem. Phys.* **7**, 61.
 Van Vleck, J. H. 1948 *Phys. Rev.* **74**, 1168.
 Whitmer, C. A., Weidner, R. T., Hsiang, J. S. & Weiss, P. R. 1948 *Phys. Rev.* **74**, 1478.

Paramagnetic resonance spectra of five chromic sulphate alums at low temperatures

BY B. BLEANEY, F.R.S.

Clarendon Laboratory, University of Oxford

(Received 31 May 1950—Revised 24 July 1950)

The paramagnetic resonance spectra of five chromic sulphate alums have been examined at temperatures down to 20° K and wave-lengths 3 and 10 cm. Their behaviour is markedly different as follows. (a) Rubidium, caesium and methylamine alums show the normal spectrum associated with a trigonal splitting at all temperatures, but the splitting decreases as the temperature is lowered, becoming constant below 90° K. The change in the splitting is most marked for rubidium, but substantially zero for methylamine alum. (b) Ammonium alum shows the normal spectrum at high temperatures, with a steadily decreasing splitting down to 80° K; here a crystallographic transition takes place, and the spectrum becomes abnormal, with two splittings. (c) Potassium alum behaves like ammonium alum down to 160° K, below which an anomalous spectrum appears with lines whose intensity increases as the temperature falls. The two splittings in this region are difficult to reconcile with previous measurements of the specific heat of the spin system by other methods.

1. INTRODUCTION

Potassium chromic alum has been extensively used for the investigation of the region below 1° K by means of the technique of adiabatic demagnetization. Its susceptibility agrees closely with the 'spin-only' value, and follows Curie's law with a divergence amounting only to a few tenths of 1 % at 1° K. The ground state of the Cr^{+++} ion is effectively ^4S , the orbital momentum being quenched by the crystalline electric field. The latter resolves the spin quadruplet into two doublets, with a separation of about 0.2 cm^{-1} , according to the theory of Van Vleck (1939); the theoretical calculation of Hebb & Purcell (1937) of the corresponding anomaly in the entropy and specific heat below 1° K are compared with the experimental values in the following paper. The specific heat of the spin system has also been determined by the method of paramagnetic relaxation. These methods offer only indirect evidence as to the splitting of the spin quadruplet, however, whereas paramagnetic

resonance makes possible the direct investigation of the energy levels of the paramagnetic ion. The nature of the spectrum can be predicted from the work of Van Vleck (1939) and Broer (1942), and has been confirmed by measurements at room temperature by Bagguley & Griffiths (1947, 1950), Weiss, Whitmer, Torrey & Hsiang (1947), Whitmer, Weidner, Hsiang & Weiss (1948) and Halliday & Wheatley (1948). The splittings deduced were 0.12 cm.^{-1} for the potassium alum, and 0.135 cm.^{-1} for the ammonium alum. The considerable discrepancy between these values and previous results by other methods at low temperatures led Bleaney & Penrose (1948) to investigate the paramagnetic resonance spectrum at temperatures down to 20° K. It was found that the splitting changed markedly with temperature in both these alums, and the spectrum was quite different from the normal at the lowest temperatures. This paper presents the results of further work on these alums, and on rubidium, caesium and methylamine chromic alums which have been similarly studied for comparison (Bleaney 1949).

The analysis of these spectra rests on premises which are fully discussed in the preceding paper by Bagguley & Griffiths (1950; henceforward referred to as I). General reference should be made to this paper, as much of the preliminary discussion is presented below in condensed form.

2. PRELIMINARY DISCUSSION

The crystallographic structure of the alums has been very fully investigated by Lipson & Beevers (1935) and Lipson (1935). The unit cell, which is cubic, contains four inequivalent chromium ions, each subject to a crystalline electric field of predominantly cubic symmetry. In some cases (see Van Vleck 1939) there is superimposed on this a small trigonal component whose axis of symmetry is directed along one of the body diagonals of the unit cube, the diagonals being different for each of the four ions. This trigonal component causes a splitting of the otherwise quadruply degenerate ^4S state into two doublets. Furthermore, when a magnetic field is applied, the behaviour of the energy levels depends on the angle θ which the field makes with the trigonal axis. Thus the paramagnetic resonance spectrum which arises from transitions between these levels also depends on θ , and is specified by the value of the parameter $c = \cos^2 \theta$. In general, the four different ions give four superimposed spectra, but simple cases arise when the magnetic field is directed along the axis of unit cube (001), or a body diagonal (111) axis. In the former case, the four ions give identical spectra; in the latter the magnetic field is parallel to the trigonal axis of one ion, for which case the behaviour of the energy levels is particularly simple. The levels diverge linearly as the strength of the field is increased (see figure 2 of I), and lie respectively at

$$W_1 = -a + \left(\frac{3}{2}\right)g\beta H,$$

$$W_2 = \left(\frac{1}{2}\right)g\beta H,$$

$$W_3 = -\left(\frac{1}{2}\right)g\beta H,$$

$$W_4 = -a - \left(\frac{3}{2}\right)g\beta H.$$

Here g is the 'spectroscopic splitting factor' in the nomenclature of Kittel (1949), and β the Bohr magneton. Since the allowed transitions involve changes only of

± 1 in the magnetic quantum number, the spectrum consists of a central line at $h\nu = g\beta H$, flanked by two equidistant satellites at $g\beta H + a$. Thus the central line gives the value of g , which will be close to the value 2 for a free spin, and the separation of the satellites gives the splitting parameter a . If the spectrum is observed, as is usually the case, at constant frequency and varying H , the spectrum is similar, the satellites lying at fields above and below that of the central line by $(a/g\beta)$, which equals $21.4a/g$ (kG) if a is measured in wave numbers.

Superimposed on this spectrum is that due to the other three ions in unit cell, for each of which $c = \cos^2 \theta = \frac{1}{9}$. The energy levels for this case, and therefore the spectrum, are fixed once the parameter a is known (see figure 2 of I). The same is true when the magnetic field is applied in the (001) or any other direction. It is thus possible to check the correctness of the initial assumption of a trigonal component of the crystalline electric field. To do so, a 'transition diagram' is plotted showing the fields at which absorption peaks occur, at various frequencies, together with the calculated variation, for the assumed value of a . In some cases it turns out that the observed transition diagram cannot be fitted by the calculated diagram for any value of the splitting. The latter can be found then only by plotting the absorption peaks back into zero magnetic field, when the splitting must be equal to the quantum corresponding to the radio-frequency field. This requires a wide wave-length coverage, but the observations provide a more detailed check of the theory than measurements confined to a single wave-length.

3. APPARATUS

The apparatus used at 3 cm. wave-length, and the method in general, is the same as that described previously (Bleaney, Penrose & Plumpton 1949). In these experiments, however, an almost continuous wave-length coverage between 3 and 10 cm. was required, and this was achieved partly by modifications of the 3 cm. apparatus, partly with a separate apparatus for the longer wave-lengths.

Between 3.0 and 4.8 cm. wave-length, the 3 cm. resonator was tuned to the required wave-length mainly by insertion of disks of dielectric (quartz or paraffin wax). In addition, the length of the resonator (which is split in the middle by a ground joint) was varied by the use of lower halves of different lengths. This latter method becomes ineffective near 3.8 cm., the cut-off wave-length of the empty cylindrical resonator. When partly filled by solid dielectric, the cavity could be tuned to resonance up to 4.8 cm.

In the region 5.6 to 11 cm., a quarter-wave coaxial line resonator was used. A series of such resonators were made, tuned to wave-lengths differing by 0.8 cm., which could be inserted inside a common jacket fitted to the apparatus by a ground joint. The paramagnetic crystal lay on the bottom (closed end of the coaxial line), and the open end faced upwards. Centimetre-wave power was fed to and from the cavity by probes at the open end, and intermediate wave-lengths could be obtained by tuning with a thin quartz tube sliding over the centre conductor of the coaxial line, and driven by a micrometer head located outside the cold part of the apparatus. The assembled apparatus is surrounded by a Dewar vessel containing the refrigerant. A complete description of the apparatus will be published elsewhere.

4. RESULTS

Since the behaviour of the spectra of the various chrome alums is very different at low temperatures, it is convenient to group the results accordingly. Qualitatively, the alums studied fall into three classes:

(1) Rubidium, caesium and methylamine alums, which exhibit the 'normal' spectrum at all temperatures, though with a splitting which varies with the temperature.

(2) Ammonium alum, which has a crystallographic transition near 80° K, below which the spectrum is quite different from the normal.

(3) Potassium alum, which has an abnormal spectrum at low temperatures but no sudden transition.

The less common alums of class (1) are considered first because their spectra conform to the usual theory, and it is more convenient to review the abnormal spectra later.

Class (1) Rubidium, caesium and methylamine alums

Some crystals of the caesium alum became available first through the courtesy of Dr D. M. S. Bagguley, and this salt was studied extensively to examine how closely its spectrum agreed with that calculated on the normal theory. Measurements were made at 90 and 20° K over a wide range of wave-lengths between 10 and 3 cm., with the magnetic field directed along each of the three directions (111), (110) and (100). No differences were observed in the positions of the absorption peaks at the two temperatures, and the results were therefore plotted together in a 'transition diagram' for each direction. Diagrams for the (111) and (100) directions are shown in figures 1 and 2. The calculated positions of the absorption peaks (from work of Professor M. H. L. Pryce) are shown as continuous or broken lines, taking $\alpha = 0.133 \pm 0.002 \text{ cm.}^{-1}$ as the value of the one unknown parameter. This value is obtained by taking an average of the separation of the measurements at different frequencies in magnetic field of the three peaks lying on these straight continuous lines in figure 1. It will be seen that discrepancies between the experimental and the calculated points are within the experimental error; in this connexion it should be remembered that the peak of an absorption line is somewhat displaced when it is overlapped by the tail of a neighbouring peak. No correction has been applied for this in plotting the experimental points.

When the magnetic field is directed along the (111) axis, the position of the central peak should correspond exactly to the equation for a single spin $H = 21.4\bar{\nu}/g$ (kG). If the spin were completely free, g should equal 2.00, and the calculated lines in the transition diagrams are drawn on this assumption. Since the orbital quenching is not absolutely complete, one would expect g to be a little less than 2. Reference to figure 1 shows that the experimental points lie a little higher than the central line, suggesting that g is appreciably less than 2. The central peak is, however, slightly displaced by a nearby peak on the high-field side (see figure 3), though insufficiently to account for all the displacement. One may take $g = 1.98 \pm 0.02$, on a conservative estimate of the errors due to these effects and in calibration of the magnetic field. Whitmer, Weidner, Hsiang & Weiss (1948) give 1.99 for undiluted ammonium and potassium chrome alums; their measurements were made at a single frequency,

however, and the cavity is not retuned to allow for anomalous dispersion of the magnetic susceptibility. It seems doubtful therefore whether a higher accuracy can be claimed for their results.

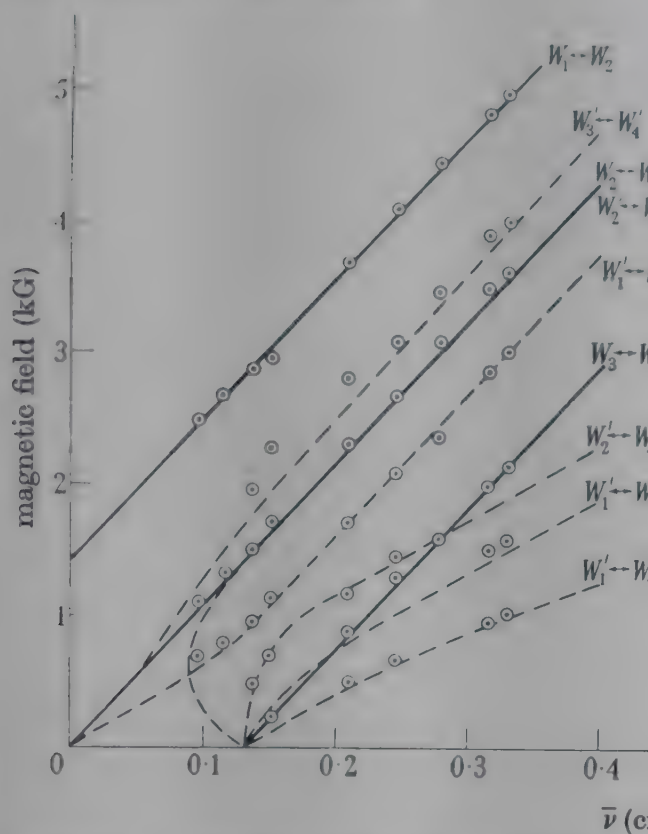


FIGURE 1. Caesium chrome alum. Transition diagram for magnetic field along (111) axis. Calculated for $a = 0.133 \text{ cm.}^{-1}$: — $c = \cos^2 \theta = 1$; ---- $c = \frac{1}{3}$. The transitions are labelled to correspond with the energy levels shown in figure 2 (I).

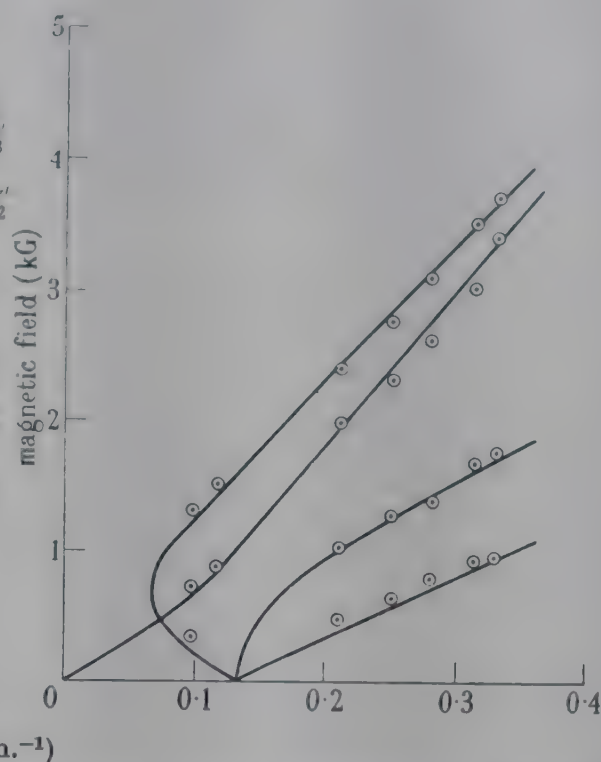


FIGURE 2. Caesium chrome alum. Transition diagram for magnetic field along (001) axis ($c = \frac{1}{3}$). — calculated for $a = 0.133 \text{ cm.}^{-1}$.

Typical absorption spectra for caesium chrome alum at 20°K with the magnetic field directed along the (111) axis are shown in figures 3 and 4, for wave-lengths of 3.16 cm. ($\bar{\nu} = 0.316 \text{ cm.}^{-1}$) and 8.60 cm. ($\bar{\nu} = 0.116 \text{ cm.}^{-1}$) respectively. In the latter case the frequency at which the measurement was made is rather lower than the initial splitting of the energy levels (0.133 cm.^{-1}), but there is still considerable absorption in zero magnetic field owing to the finite width of the levels and their complicated behaviour in low magnetic fields. (Reference to figure 1 shows that there are a number of possible transitions which converge to the value 0.133 cm.^{-1} as the magnetic field approaches zero.) The thick vertical lines show the calculated positions of the transitions (assuming $g = 2$ exactly) for $a = 0.133 \text{ cm.}^{-1}$, and reference to figure 1 identifies the energy levels to which they belong. In figure 3 the transitions at fields of 0.93 , 1.49 and 1.89 kG with small intensities correspond to jumps of 2 and 3 in the magnetic quantum numbers. These are allowed transitions in the region where the effect of the magnetic field is comparable with the initial splitting, since the strong-field quantum numbers are not 'good' quantum numbers while the ion is changing its axis of quantization from the trigonal axis of the crystalline field to the external magnetic field.

Before proceeding to the question of intensity in these spectra some discussion of the line width is required. The spin-lattice relaxation time in chrome alum at 20°K is known (Gorter 1947) to be longer than 10^{-6}sec. and its contribution to the line width would be less than 1 G. The line width must therefore be due to spin-spin interaction, and in these dilute salts exchange effects should be so small that only magnetic dipole-dipole interaction need be considered. Explicit formulae for this have not been developed for a case in which there is an initial splitting of the spin quadruplet, but in strong fields one may use as a first approximation the formulae of Van Vleck (1948) with the following modification. The frequency with which the

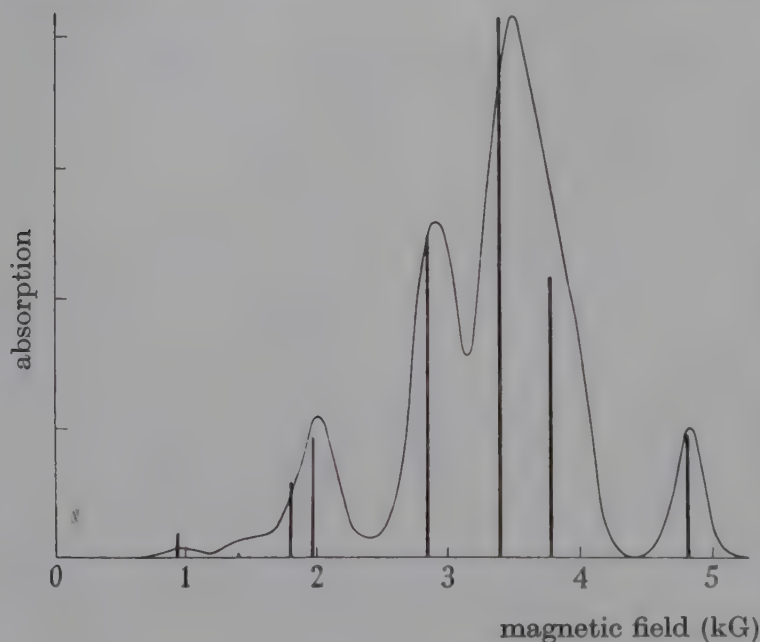


FIGURE 3. Caesium chrome alum. Spectrum at $\bar{\nu} = 0.316\text{ cm.}^{-1}$, magnetic field along (111) axis.

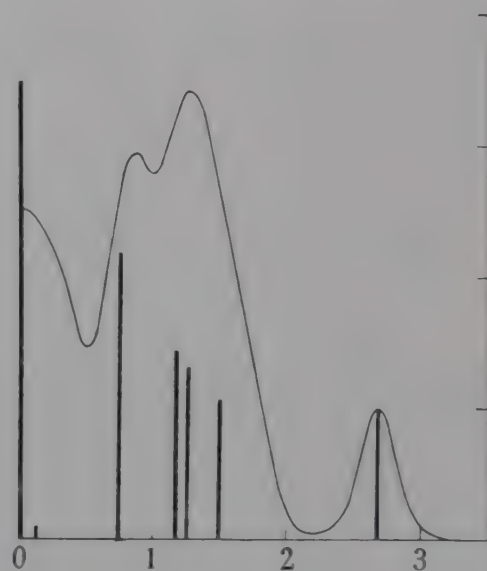


FIGURE 4. Caesium chrome alum. Spectrum at $\bar{\nu} = 0.116\text{ cm.}^{-1}$, magnetic field along (111) axis.

The calculated spectra are shown by the thick vertical lines.

ions precess in the magnetic field depends on the angle θ which the latter makes with the trigonal axis. For the (111) direction the two types of ion, for which $c = \cos^2\theta = 1$ and $\frac{1}{9}$ respectively, precess with different frequencies, and are unable to exchange quanta with one another. In calculating the line width these must be counted as 'non-identical spins' in Van Vleck's nomenclature, and their contribution to the mean square field is smaller by a factor $\frac{4}{9}$ than that due to 'identical spins'. Proceeding in this way, one finds that the root mean square width

$$H_i = 118\text{ G for ions } c = 1,$$

$$H_i = 152\text{ G for ions } c = \frac{1}{9}.$$

In figure 5 the high-field peak of figure 3 is replotted on a large scale showing the experimental points. The continuous line on which these points all lie within the experimental error represent a Gaussian shape, $\exp(-\Delta H^2/2H_i^2)$, where H_i has the calculated value 118 G. The theory of Van Vleck shows that for a simple cubic

lattice lines observed with the magnetic field in the (111) direction should not depart greatly from the Gaussian shape, and the experimental result is in good agreement with this.



FIGURE 5. Caesium chrome alum. Detail of high-field peak of figure 3.
 ○ experimental points; — theoretical curve assuming Gaussian shape.

If the assumption is made that the line shape is similar for the transitions belonging to the $c = \frac{1}{9}$ ions, but with the greater width given above, the theoretical spectrum can be constructed using the transition intensities calculated by G. A. Wheatley. This has been carried out for this salt at room temperature by Bagguley & Griffiths (I), and we shall content ourselves here with drawing the heights of the thick vertical lines in figures 3 and 4 to correspond with the peak intensity of the line. Since only relative intensity measurements have been made, the calculated peak intensities have been adjusted to fit one line in each figure, the central line in figure 3 and the high-field peak in figure 4. In the former case the agreement with experiment is good except for the partly forbidden lines in low fields. This may be due to the fact that, though the r.f. field is perpendicular to the external field, the ions are precessing about an intermediate axis owing to the effect of the crystalline field. The discrepancy is very marked for the absorption in zero field in figure 4. If the intensity of the latter is taken to be the same as if the frequency of observation were exactly equal to the splitting $\alpha = 0.133 \text{ cm.}^{-1}$, it should be about 0.46 in the units of figure 4.

In arriving at this figure, allowance has been made for the fact that the r.f. field is not perpendicular to the axis of precession of each ion; but the line widths have been assumed the same as in the strong field case. It is probably more accurate to treat all ions as identical in zero field, in which case the theoretical intensity would be

about 0.35. Pryce & Stevens (1950) have pointed out that the line width is also dependent on the orientation of the r.f. magnetic field, in a case such as this. Thus these intensities are rather uncertain until a proper calculation of the line widths has been carried through, for this case, and for the partly-forbidden low-field lines in figure 3.

Similarly extensive measurements were not carried out on rubidium and methylamine alums, but it was verified that the spectra correspond to those shown for caesium alum. The splitting varies with temperature, however, between 90 and 290° K, and measurements were made also at the dry ice point (193° K). The splittings for the three alums are shown in table 1.

TABLE 1. SPLITTINGS (CM.⁻¹) IN VARIOUS CHROMIC ALUMS AT LOW TEMPERATURES

temperature (° K)	ammonium	potassium	rubidium	caesium	methylamine
290	0.135	0.12	0.165	0.145	0.165
193	0.085	0.055	0.126	0.134	—
90	0.035 transition point	{ 0.26 0.15 ± 0.01	0.108 ± 0.002	0.133 ± 0.002	0.170 ± 0.003
80	{ 0.314 ± 0.003 0.242 ± 0.003	—	—	—	—
20	{ 0.317 ± 0.003 0.240 ± 0.003	{ 0.270 ± 0.003 0.15 ± 0.01	0.108 ± 0.002	0.133 ± 0.002	0.170 ± 0.003

* The values at room temperature are taken from the preceding paper by Bagguley and Griffiths.

The only independent determination of the splitting is that for the caesium alum at 2° K, obtained by Benzie & Cooke (1950) from paramagnetic relaxation measurements. This value is 0.138 ± 0.002 cm.⁻¹, which is just outside the experimental error from agreement. It seems unlikely that the splitting should change between 20 and 2° K, and it may be that the contribution to the specific heat from spin-spin interaction is greater than that due to magnetic dipole interaction, as assumed by Benzie & Cooke. If it were greater by a factor 2, the discrepancy would disappear. Such a factor would not be unreasonable in view of the contributions from exchange to the spin specific heat in the copper tutton salts, which are not greatly different in their degree of magnetic dilution.

Class (2) Ammonium chrome alum

Paramagnetic resonance experiments on the alums at low temperatures were initiated with this salt because of the enormous discrepancy between the room temperature splitting (0.135 cm.⁻¹) found by Bagguley & Griffiths (1950), and the value of 0.34 cm.⁻¹ at 77° K deduced from the measurements of Starr (1941). A preliminary account of the behaviour of the spectrum has been published by Bleaney & Penrose (1948), but a further study has revealed one important new feature and a number of subsidiary points. It is therefore preferable to give a report of the whole work.

At room temperature Bagguley & Griffiths have found the spectrum of this alum to correspond closely to the normal with a splitting of 0.13₅ cm.⁻¹. As the temperature

is lowered this splitting diminishes at the rate of 0.05 cm.^{-1} per 100° K (see table 1 and Bleaney & Penrose 1948), becoming so small that no resolved side-peaks (except that corresponding to a jump of 2 in the magnetic quantum number) can be observed at 90° K , though the width is greater than would be expected for a single line. From the experimental value of the mean square width H_i^2 , the splitting can be estimated as 0.03_5 cm.^{-1} , assuming that the unresolved component lines have the same relative separations and intensities as in the normal spectrum.

At 81° K a transition takes place, which is not instantaneous but occupies a time of a few minutes. Experimentally the transition is manifest through the large shift in the resonance frequency of the cavity which occurs, which must be associated with a discontinuity in the dielectric constant of the salt. The detuning is so rapid that the spectrum cannot be observed during the transition, but subsequently the spectrum is found to be quite different from that above the transition point, and it cannot be fitted to a normal spectrum for any value for the splitting. This spectrum remains substantially unaltered on cooling to liquid hydrogen temperatures. On rewarming, the reverse transition does not occur until 105° K , when the normal spectrum returns and the splitting at higher temperatures is the same as that observed during cooling. Similar behaviour is found on subsequent temperature cycles, though the temperature hysteresis in the transition tends to diminish slightly.

The two spectra, just above and just below the transition point, are shown in figure 6, for the magnetic field parallel to the (111) axis. The presence of absorption in zero magnetic field in the second spectrum shows that the frequency of measurement is almost identical with a splitting. An accurate value is obtained by measurements at slightly higher frequencies, when an absorption peak is observed at very low fields, which are plotted against the frequency and extrapolated back to zero field. Hence the splitting is found to be $0.314 \pm 0.003 \text{ cm.}^{-1}$ at 80° K , increasing slightly to 0.317 ± 0.003^{-1} at 20° K .

In a preliminary report of this discovery, Bleaney & Penrose (1948) assumed that only one splitting could be present since the Kramers degeneracy cannot be lifted by a crystalline electric field. A detailed plot against frequency of the magnetic fields at which peaks occur (figure 7) reveals a second splitting equal to $0.242 \pm 0.003 \text{ cm.}^{-1}$ at 20° K .

Discussion

The transition is undoubtedly the same as that found by Kraus & Nutting (1941), who observed that the crystal became opaque as if 'shattered' into innumerable small crystals. In our apparatus the crystal could be viewed only at room temperature; a thin slice, moderately transparent before cooling, becomes opaque and brittle afterwards. The paramagnetic resonance spectrum indicates that below the transition point this alum must have a different crystal structure probably containing two inequivalent chromium ions subject to different crystalline fields. The stress of transition shatters the crystal, which will not, of course, reform as an entire crystal on returning to the normal structure, though the micro-crystals must have parallel axes.

At low temperatures, the splittings observed are sensibly the same for different crystals, though the transition diagram shows some scatter in the high-field peaks. Mixed crystals with ammonium aluminium alum, with Cr:Al about 1:3, gave a similar low-temperature spectrum, but with the side-peaks split up into many subsidiary peaks. A more dilute specimen (Cr:Al about 1:30), however, gave a normal spectrum with the side-peaks much broader (and correspondingly weaker) than the central line. Since the broadness of the side-peaks increased with their distance from

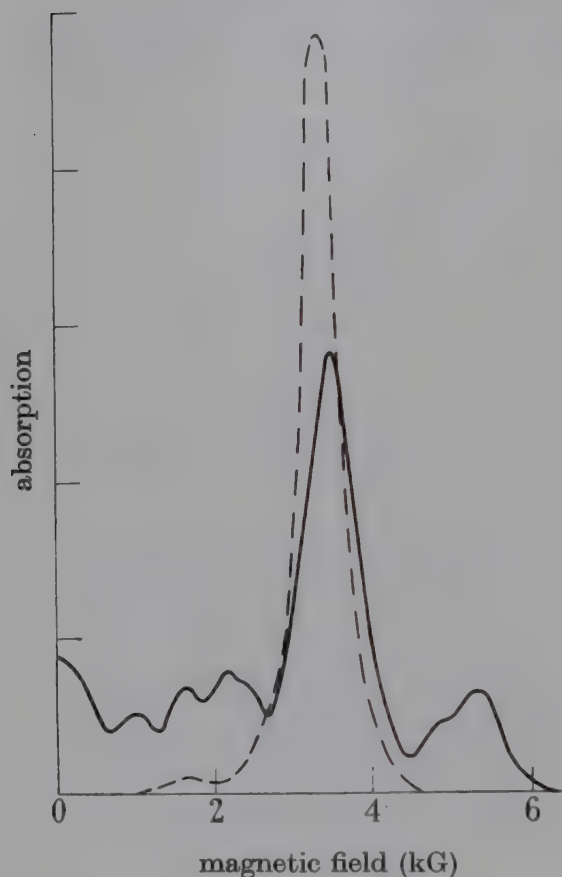


FIGURE 6. Ammonium chrome alum. Spectra for magnetic field along (111) axis near transition point (81°K). ---- Spectrum just above transition point; — spectrum just below transition point.

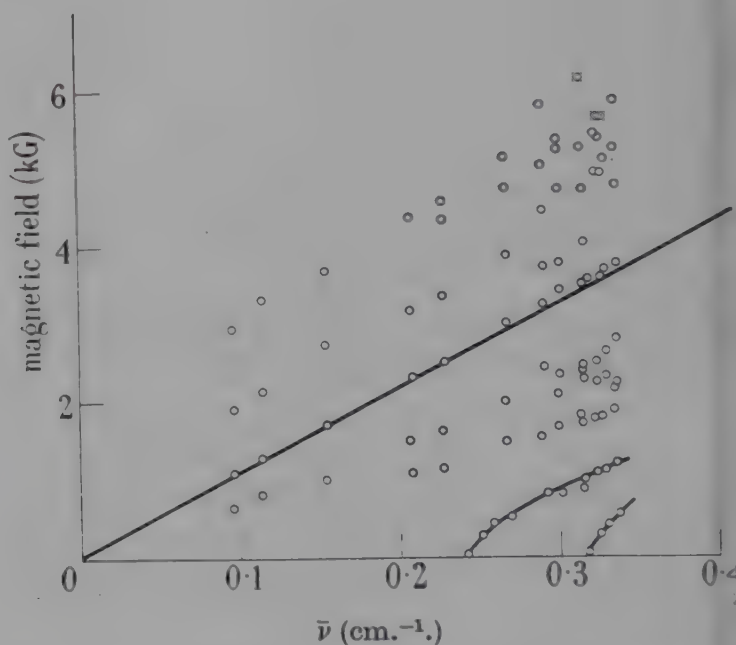


FIGURE 7. Ammonium chrome alum. Transition diagram for magnetic field along (111) axis showing two splittings in zero field.

the central peak (whose position is independent of the splitting), it seems that the splitting is not quite constant within the mixed crystal. This is probably due to local inhomogeneities associated with the random distribution of Cr ions, whose desire to partake of a different crystal structure is not shared by the more prevalent Al ions.

It is difficult to obtain a precise value for the relative numbers of chromium ions with the two splittings, since this involves a comparison of intensities and line widths at different frequencies. One may estimate, however, that they do not differ by more than a factor 2. This accords with a specific heat determination at 2°K by Benzie & Cooke (1950), which corresponds to a mean splitting of 0.27 cm.^{-1} .

Class (3) Potassium chrome alum

At high temperatures the behaviour of this alum is similar to that of the ammonium alum; the spectrum is normal, and the splitting falls linearly with temperature from 0.12 cm.^{-1} at 290° K to 0.03_5 cm.^{-1} at 160° K . The latter value is obtained by analysis of the mean square line width as for the ammonium alum. At this temperature, however, there is no sudden transition, but instead side-peaks appear as the temperature is lowered further. The position of these side-peaks is nearly independent of temperature, but their intensity relative to the central peak grows steadily with decreasing temperature. The spectrum bears a superficial resemblance to the normal type for a splitting of about 0.17 cm.^{-1} (Bleaney & Penrose 1948; Halliday & Wheatley 1948). The intensity of the side-peaks is very much less than normal, and careful investigation over a wide range of wave-lengths reveals that there are two splittings in this salt as in the ammonium alum. The values of these splittings are given in table 1, and were obtained by plotting resonance peaks back into zero magnetic field. In the case of the smaller splitting (0.15 cm.^{-1}) the peaks were so broad that the accuracy of this method was only $\pm 0.01 \text{ cm.}^{-1}$. Absorption in zero magnetic field at frequencies equal to the two splittings is clearly revealed in figure 8. Such absorption is relatively much less strong than in the corresponding case for the caesium alum (figure 4).

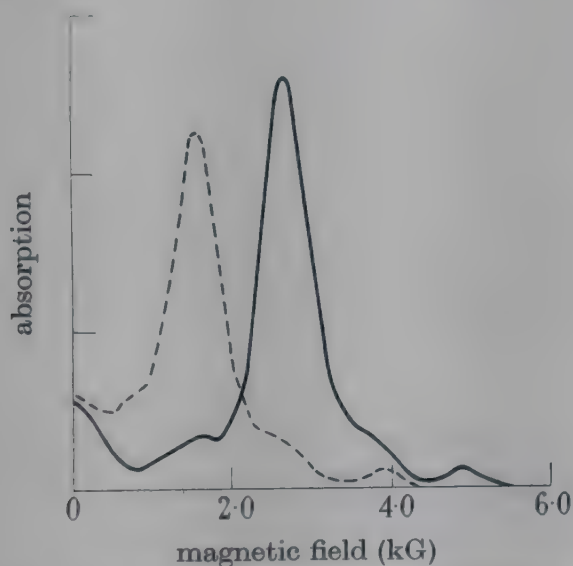


FIGURE 8. Potassium chrome alum. Spectra for magnetic field along (111) axis.
— $\bar{\nu} = 0.270 \text{ cm.}^{-1}$; ---- $\bar{\nu} = 0.15 \text{ cm.}^{-1}$.

In many respects the behaviour of this alum is similar to that of the ammonium alum, except that no sudden transition occurs. The gradual change in intensity of the side-peaks at low temperatures suggests that the potassium alum is trying to change to the two-splitting modification, but the change is incomplete. Only a fraction of the ions are subject to the abnormal crystalline field, a fraction which increases as the temperature falls.

It seems impossible to reconcile this suggestion with data on the specific heat associated with the spin system. Paramagnetic relaxation measurements of Gorter, Dijkstra & van Paemel (1942) and Broer (1947) show that this specific heat corresponds to the same Stark splitting of the spin levels at 77 and 90° K as is found in the helium

region and below. Moreover, this splitting (assumed the same for all ions) is 0.17 cm.^{-1} . The two splittings shown in this paper to exist are 0.15 and 0.27 cm.^{-1} , which can only be fitted quantitatively with the specific heat tail by assigning (*ad hoc*) a suitable fraction of the ions to each splitting. If all of the ions have one or the other of these two splittings, the fraction with 0.27 cm.^{-1} must be only 15 %; this conflicts with the spectrum intensities, which indicate roughly equal fractions. If then we take 30 % of the ions with $a = 0.15 \text{ cm.}^{-1}$, 30 % with $a = 0.27 \text{ cm.}^{-1}$ and the remaining 40 % as retaining the very small $a = 0.035 \text{ cm.}^{-1}$ (which would fit both the intensities and the specific heat tail), the specific heat anomaly to be expected below 1° K is vastly different from the experimental curve (Bleaney 1950). In addition, no simple distribution of splittings can explain the fact that the plateau in the entropy curve found by de Klerk, Steenland & Gorter (1949) falls below $R \log_e 2$, since Kramers's theorem requires that no electric field can lift the twofold degeneracy of the system.

The possibility of a hyperfine structure seems to provide a way out of this latter difficulty, especially as Ingram & Scovil (unpublished) have found an extensive *hfs* in $\text{V}(\text{NH}_4)_2(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$, where the vanadium is iso-electronic with trivalent chromium. It is obvious at once, however, that our paramagnetic resonance spectrum cannot be explained in this way, since the only odd isotope of chromium naturally occurring is ^{53}Cr , of abundance 9 %. Lines split up by interaction with the spin of this nucleus would have therefore a relative intensity of only a few per cent. Attempts to detect an *hfs* by using crystal highly diluted with potassium aluminium alum ($\text{Cr}:\text{Al} \approx 1:1000$) have been thwarted because only a single narrow line could be seen corresponding to the central transition, whose position is independent of the Stark splitting. Presumably the latter is non-uniform in the mixed crystal (as in the diluted ammonium alum), and the side-peaks are therefore smeared out, becoming too weak to be visible. This difficulty with diluted salts has been found only with the alums, showing that the chromium and aluminium alums are not truly isomorphous.

A minor point of interest is that the anomalous behaviour with falling temperature of the potassium has the same characteristics as were observed in the optical spectrum by Spedding & Nutting (1934). The spectra at 85 and 14° K were notably different, but at intermediate temperatures both spectra were present with varying relative intensity.

5. GENERAL DISCUSSION

A detailed analysis of the crystal structure of the chrome alums has not been carried out, but Lipson (1935) has shown that there are significant differences in the structure of a number of aluminium alums. These are related to the size of the monovalent cation, and Van Vleck (1939) has assumed that there are similar differences in the chrome alums, though the correlation has been verified only for the potassium alum. Lipson places the caesium and methylamine alums in structural class β , and ammonium, potassium and rubidium in class α , though rubidium can also crystallize in the β form. This division corresponds to our classification, except for rubidium, which with its temperature-dependent splitting could be regarded as an intermediate case, if the classification is stretched.

Quantitatively, it seems doubtful whether the differences in structure observed by Lipson bear any relation to the splittings observed. Van Vleck (1939, but see correction 1940) estimates that a distortion of the water molecules from a regular octahedron centred on the Cr^{+++} ion amounting only to 10^{-10} cm. would be sufficient to account for the observed splittings. One would therefore expect considerable differences in splitting between, for example, the ammonium alum, where distortion can be detected by X-ray methods whose precision is about 30×10^{-10} cm., and the caesium alum, where it cannot. The X-ray measurements were made at room temperature, where the splitting is almost identical for these two alums. One may conclude that probably the most remarkable feature of the X-ray measurements is that the predicted trigonal distortion is confirmed at room temperature for all the alums whose paramagnetic resonance spectrum has been observed at that temperature!

When the sudden transition in ammonium chrome alum was discovered by Bleaney & Penrose (1948) it was assumed that this corresponded to the jump in the dielectric constant found by Guillien (1939). Later work of this author (Guillien 1941, 1943) shows, however, that in ammonium aluminium alum the dielectric constant corresponds to a Debye-type relaxation which gives a sharp anomaly when observed at constant frequency and varying temperature. Though some of the alums appear to show abnormal dielectric loss in the temperature region around 150°K where Guillien's relaxation time (extrapolated) would be of the same order as the period of the r.f. oscillations, there seems to be no obvious correlation with the paramagnetic resonance spectra.

The author wishes to record his debt to the late Dr R. P. Penrose and Miss B. I. Plumpton, who made most of the measurements described above; and to Drs Bagguley and Griffiths for generous help and discussion.

REFERENCES

- Bagguley, D. M. S. & Griffiths, J. H. E. 1947 *Nature*, **160**, 532.
 Bagguley, D. M. S. & Griffiths, J. H. E. 1950 *Proc. Roy. Soc. A*, **204**, 188.
 Benzie, R. J. & Cooke, A. H. 1950 *Proc. Phys. Soc. A*, **63**, 213.
 Bleaney, B. 1949 *Phys. Rev.* **75**, 1962.
 Bleaney, B. 1950 *Proc. Roy. Soc. A*, **204**, 216.
 Bleaney, B. & Penrose, R. P. 1948 *Proc. Phys. Soc.* **60**, 395.
 Bleaney, B., Penrose, R. P. & Plumpton, B. I. 1949 *Proc. Roy. Soc. A*, **198**, 406.
 Broer, L. J. F. 1942 *Physica*, **9**, 547.
 Broer, L. J. F. 1947 *Physica*, **13**, 353.
 de Klerk, D., Steenland, M. J. & Gorter, C. J. 1949 *Physica*, **15**, 649.
 Gorter, C. J. 1947 *Theory of paramagnetic relaxation*. Amsterdam and New York: Elsevier Publishing Co. Ltd.
 Gorter, C. J., Dijkstra, L. J. & van Paemel, O. 1942 *Physica*, **9**, 673.
 Guillien, R. 1939 *C.R. Acad. Sci., Paris*, **209**, 21.
 Guillien, R. 1941 *C.R. Acad. Sci., Paris*, **213**, 991.
 Guillien, R. 1943 *C.R. Acad. Sci., Paris*, **217**, 443.
 Halliday, D. & Wheatley, J. 1948 *Phys. Rev.* **74**, 1478.
 Hebb, M. H. & Purcell, E. M. 1937 *J. Chem. Phys.* **5**, 338.
 Kittel, C. 1949 *Phys. Rev.* **76**, 743.
 Kraus, D. L. & Nutting, G. C. 1941 *J. Chem. Phys.* **9**, 133.

- Lipson, H. 1935 *Proc. Roy. Soc. A*, **151**, 347.
Lipson, H. & Beevers, C. A. 1935 *Proc. Roy. Soc. A*, **148**, 664.
Pryce, M. H. L. & Stevens, K. W. H. 1950 *Proc. Phys. Soc. A*, **63**, 36.
Spedding, F. H. & Nutting, G. C. 1934 *J. Chem. Phys.* **2**, 421.
Starr, C. 1941 *Phys. Rev.* **60**, 241.
Van Vleck, J. H. 1939 *J. Chem. Phys.* **7**, 61.
Van Vleck, J. H. 1940 *Phys. Rev.* **57**, 426.
Van Vleck, J. H. 1948 *Phys. Rev.* **74**, 1168.
Weiss, P. A., Whitmer, C. A., Torrey, H. C. & Hsiang, J. S. 1947 *Phys. Rev.* **72**, 975.
Whitmer, C. A., Weidner, R. T., Hsiang, J. S. & Weiss, P. A. 1948 *Phys. Rev.* **74**, 1478.
-

Thermal properties of potassium chromic alum between 0.05 and 1°K

BY B. BLEANEY, F.R.S., *Clarendon Laboratory, University of Oxford*

(Received 31 May 1950—Revised 24 July 1950)

The entropy, specific heat and magnetic temperature (reciprocal of the susceptibility) of potassium chromic alum are measured as functions of the absolute temperature between 0.05 and 1° K. Their interpretation in the light of paramagnetic resonance measurements (preceding paper) is discussed.

1. INTRODUCTION

In 1938–9 the author carried out a series of measurements on the thermal properties of potassium chromic alum below 1° K, no detailed account of which has been published, though the results have been quoted on a number of occasions (Van Vleck 1940; Cooke 1949). The method of paramagnetic resonance gives an independent determination of the splitting parameters involved in the interpretation of the low temperature measurements, and it seemed of interest to discuss the latter in the light of the results reported in the preceding paper (Bleaney 1950). Moreover, de Klerk, Steenland & Gorter (1949) have recently published a far-reaching investigation of the thermal and magnetic properties of this substance in the region below 0.05° K, to which the author's results are complementary. The latter are presented below, without further excuse for their publication.

2. THE EXPERIMENTS

A very full account of the experimental technique used in these measurements has already been published by Hull (1947), so that only a brief outline is required here.

The sample consisted of an ellipsoid of powdered alum compressed to within 3 or 4 % of the crystalline density, suspended in a chamber surrounded by liquid helium at 1° K. The mode of suspension is shown in figure 1 of Hull (1947). The ratio, length to diameter, of the ellipsoid was 4:1 and its weight was 4.06 g.

The magnetic temperature was measured by a mutual inductance method, the reversal of a known current in the primary coil producing a ballistic throw in a galvanometer connected to a secondary coil. The magnetic temperature was reduced to that appropriate to a sphere, T^* , the correction constant being 0.022°K . The magnetic thermometer was calibrated against the helium vapour pressure curve of Schmidt & Keesom (1937). Below 1.6°K deviations were observed owing to the error in this curve, which led to the recalculation and measurement of the vapour pressure curve (Bleaney & Simon 1939; Bleaney & Hull 1941). This region was therefore not used in calibrating the magnetic thermometer.

Briefly, the method consists of three parts:

(a) the establishment of an entropy $-T^*$ curve by a series of demagnetizations from known fields and temperatures of about 1°K , the entropy being calculated from the theory for a free spin of $\frac{3}{2}$;

(b) the establishment of a 'magnetic specific heat' or C^*-T^* curve, using gamma-ray heating. Typical details have been given by Hull (1947);

(c) the absolute temperature for a given T^* is calculated from the relation

$$T = \left(\frac{\partial Q}{\partial T^*} \right)_{H=0} / \left(\frac{\partial S}{\partial T^*} \right)_{H=0} = C^* / \left(\frac{\partial S}{\partial T^*} \right)_{H=0}.$$

Since with gamma-ray heating the value of C^* is known only in arbitrary units, the same is true also of T . The appropriate reduction factor is obtained by setting the limit of T/T^* , which tends to a constant value at temperatures approaching 1°K , equal to unity.

When the relation between T^* and T has been obtained, the entropy and specific heat curves may be reduced to functions of the thermodynamic temperature T . The lowest value of T^* reached was 0.0616°K , by demagnetization from a field of 13.0 kG at 1°K . The method of Kurti, Lainé, Rollin & Simon (1936) was used to test for ferromagnetic remanence, no trace of which was found. De Klerk *et al.* (1949) have since found that remanence appears only below $T^* = 0.03^\circ (0.004^\circ \text{K})$.

3. THE ENTROPY

The results of a number of demagnetizations are shown in table 1. Here H_e is the external field, H_o the combined Lorentz and demagnetizing field, and $H_e - H_o = H_i$ the actual field acting on the paramagnetic ions. All fields are expressed in kilogauss. Here T_i is the temperature at magnetization,† and T^* the temperature reached after demagnetization, corrected for any thermal drift.

To construct a smoothed curve of entropy against T^* , use is made of the fact that the quantity $\alpha = H_i T^* / T_i$ remains fairly constant over a wide range of values of H_i / T_i for this substance. The values of α shown in the last column of table 1 have been plotted against T^* and from the best curve through these points an entropy $-T^*$ curve has been constructed. Values of $(\log_e 4 - S/R)$ are given in table 2, that for 1°K in zero field being taken as $\log_e 4 - 0.008$, where the correction 0.008 represents

† T_i is measured by the magnetic thermometer, and strictly should be written T_i^* . The deviation from the absolute temperature is small (*ca.* 0.003°K) and is neglected.

the decrease in S/R due to the crystalline splitting, calculated from the inverse square law for the specific heat (see § 5 below). The contribution of the lattice vibrations at 1°K to S/R is about 10^{-4} , and has been neglected.

In the first column of table 2 is given the value of $y = H_i/T_i$; that is, the field required to reach the temperature T^* by demagnetization from 1°K exactly.

TABLE 1

T_i ($^\circ\text{K}$)	T^* ($^\circ\text{K}$)	H_s (kG)	H_o (kG)	H_i (kG)	α (kG)
0.983	0.591	1.211	0.013	1.198	0.72
0.977	0.461	1.694	0.019	1.675	0.79
0.970	0.340	2.38	0.026	2.35	0.82
0.972	0.264	3.18	0.035	3.14	0.85
0.973	0.214	3.80	0.040	3.76	0.83
0.975	0.147	5.59	0.055	5.53	0.84
0.976	0.129	6.45	0.062	6.39	0.84
0.966	0.112	7.29	0.066	7.22	0.84
0.961	0.1038	8.04	0.070	7.97	0.86
0.963	0.0845	9.39	0.08	9.31	0.82
0.970	0.0815	9.70	0.08	9.62	0.81
0.977	0.0762	10.45	0.08	10.37	0.81
0.984	0.0725	11.30	0.08	11.22	0.83
1.064	0.0754	11.77	0.08	11.69	0.83

TABLE 2

$y = H_i/T_i$	T^*	$\log_e 4 - S/R$	$y = H_i/T_i$	T^*	$\log_e 4 - S/R$
0	1.000	0.008	7.0	0.119	0.410
2.5	0.341	0.075	7.5	0.112	0.452
3.0	0.279	0.104	8.0	0.105	0.493
3.5	0.239	0.134	8.5	0.0984	0.533
4.0	0.209	0.166	9.0	0.0920	0.573
4.5	0.186	0.203	9.5	0.0871	0.611
5.0	0.167	0.242	10.0	0.0828	0.647
5.5	0.152	0.284	10.5	0.0782	0.685
6.0	0.139	0.326	11.0	0.0746	0.720
6.5	0.129	0.370			

4. THE THERMODYNAMIC TEMPERATURES

The fact that α remains practically constant from 0.3° down to 0.09° (T^*) is of great convenience in calculating the absolute temperature. In the equation

$$T = C^* / \left(\frac{\partial S}{\partial T^*} \right)_{H=0}, \tag{1}$$

the determination of dS/dT^* from the $S-T^*$ curve by a graphical method does not permit of great accuracy. From the relation

$$y = H_i/T_i = \alpha/T^*, \tag{2}$$

we have

$$\begin{aligned} \frac{\partial S}{\partial T^*} &= \frac{\partial S}{\partial y} \frac{\partial y}{\partial T^*} \\ &= \frac{\partial S}{\partial y} \left[\frac{1}{T^*} \frac{d\alpha}{dT^*} - \frac{\alpha}{T^{*2}} \right]. \end{aligned}$$

Hence

$$\frac{T}{T^*} = \frac{C^*}{\frac{\partial S}{\partial y} \left[\frac{d\alpha}{dT^*} - \frac{\alpha}{T^*} \right]} \quad (3)$$

When α is constant

$$\frac{T}{T^*} = -\frac{C^*}{\frac{\alpha}{T^*} \frac{dS}{dy}} = -\frac{C^*}{y \frac{\partial S}{\partial y}} \quad (4)$$

The formula (4) is very convenient to work with; for a given value of y , the corresponding T^* is obtained from table 2; the value of C^* is given by the gamma-ray heating experiment and dS/dy is calculated from the theoretical magnetization curve. Below 0.1°K $d\alpha/dT^*$ is not negligible and formula (3) is used, the value of $d\alpha/dT^*$ being taken from the $\alpha - T^*$ curve. Above 0.3°K α begins to vary rapidly, for equation (2) shows that it must fall to zero for $H_i = 0$ and $T^* = T_i$; $d\alpha/dT$ thus becomes of the same order as α/T . A $Q-S$ curve has therefore been constructed in this region, from which the absolute temperature is obtained by differentiating; the values so obtained are in good agreement with those obtained by using formula (3). The values of T^* for values of T throughout the range of experiments are shown in column (2) of table 3.

TABLE 3

T	T^*	C/R (exp.)	$\log_e 4 - S/R$	T	T^*	C/R (exp.)	$\log_e 4 - S/R$
1.000	1.000	0.0160	0.008	0.180	0.195	0.296	0.187
0.600	0.604	0.042	0.022	0.160	0.174	0.325	0.226
0.480	0.485	0.064	0.034	0.140	0.156	0.350	0.271
0.400	0.406	0.089	0.047	0.120	0.138	0.374	0.329
0.360	0.368	0.108	0.057	0.100	0.121	0.391	0.404
0.320	0.330	0.130	0.072	0.080	0.103	0.40	0.503
0.280	0.291	0.163	0.091	0.060	0.086	0.39	0.616
0.240	0.252	0.206	0.120	0.050	0.079	0.36	0.682
0.200	0.215	0.266	0.158	0.045	0.075	0.33	0.720

A discussion of these results and their comparison with theory of Hebb & Purcell (1937) has already been published by Cooke (1949). Figure 5 of Cooke's paper shows the fractional deviation of T^* from T , together with the theoretical curves based on three different methods of allowing for the magnetic interaction between the ions. This figure shows that either the Onsager method or the Gaussian method of Van Vleck gives a reasonable interpretation of the experimental results down to 0.05°K , while the Lorentz method is decidedly worse. It may be pointed out in this connexion that the calculation assumes a value 0.24°K for the crystalline splitting δ (see below) which is not supported by paramagnetic resonance experiments. Nevertheless, a very large change in δ would be required to bring the Lorentz curve into agreement with experiment, and the maximum ratio of T^* to T allowed by it is 1.67. De Klerk, Steenland & Gorter (1949) find that T^*/T approaches 10 at lower temperatures. The Onsager formula predicts a minimum value for T^* of 0.063° ; the Leiden measurements give a minimum value of 0.037° .

5. THE ABSOLUTE SPECIFIC HEAT

To convert the $C^* - T^*$ curve into a $C - T$ curve, the values of both T^* and dT^*/dT for various values of T were obtained from the $T^* - T$ relation. In figure 1 C/R is shown as a function of T below 0.25°K (curve A); the author's measurements extend down to 0.045°K and are joined by the broken line to meet the measured values of de Klerk *et al.* below 0.03°K (curve B).

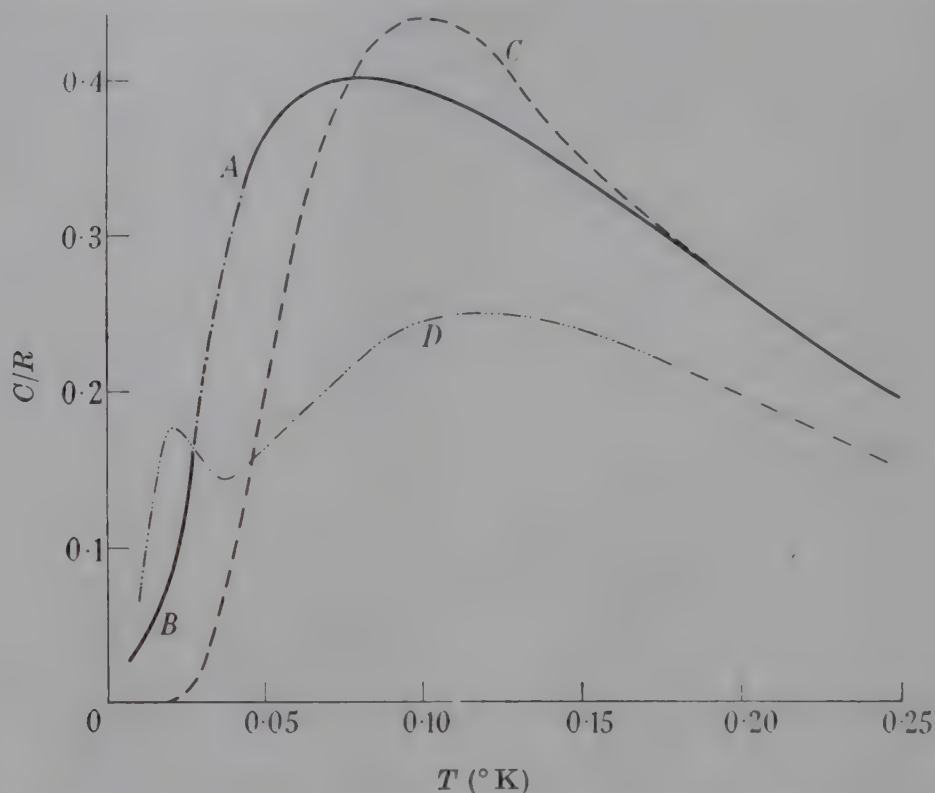


FIGURE 1. The absolute specific heat of potassium chrome alum. A , experimental curve (this paper). B , experimental curve (de Klerk *et al.*). C , calculated curve for splitting $\delta = 0.245^\circ$. D , calculated curve for mixture of splittings (see text).

Above 0.6°K the specific heat obeys the law $CT^2/R = \text{constant} = 0.0160$. From this one can obtain the splitting δ , making the usual assumption that all ions have the same splitting. From Hebb & Purcell (1937)

$$CT^2/R = 2.40\tau^2 + \frac{1}{4}\delta^2,$$

where the first term arises from the spin-spin interaction, and the second from the crystalline splitting. Taking the theoretical value of 0.0204°K for τ , one finds $\delta = 0.245 \pm 0.004^\circ \text{K} = 0.170 \pm 0.003 \text{ cm.}^{-1}$. This is in reasonable agreement with de Klerk *et al.*, who find 0.25°K , by correcting their $S - T^*$ curve to an $S - T$ curve using the formulae of Hebb & Purcell. (de Klerk *et al.* refer several times to the theoretical relation of Hebb & Purcell without stating which of the three methods of allowing for the magnetic interaction is applied; from figure 3 of their paper it appears that they intend always the Lorentz method.) In a previous set of measurements Casimir, de Klerk & Polder (1940) found $\delta = 0.26_3^\circ \text{K}$, and it has been suggested that the difference between this and the later Leiden measurements is real, and that the value of δ may vary with the origin of the salt.

The value of (T^2/R) has been determined by a number of workers using the method of paramagnetic relaxation. The values deduced for the splitting using the same assumption as above are given in table 4. In this table the first section contains the paramagnetic relaxation measurements, the second the adiabatic demagnetization measurements, and the last the splittings determined directly from the paramagnetic resonance spectrum.

TABLE 4

method	temperature (° K)	splitting		authors
		(° K)	(cm. ⁻¹)	
paramagnetic relaxation	77 and 90	0.243	0.169	Broer (1947)
	77 and 90	0.243	0.169	Gorter, Dijkstra & van Paemel (1942)
	77	0.231	0.161	Starr (1941)
	2	0.260	0.181	Casimir, Bijl & du Pré (1941)
adiabatic demagnetization	below 1°	0.270	0.188	Casimir, de Haas & de Klerk (1939)
	„	0.263	0.183	Casimir, de Klerk & Polder (1940)
	„	0.251	0.175	de Klerk, Steenland & Gorter (1949)
	„	0.245	0.170	Bleaney (this paper)
paramagnetic resonance	20°	{ 0.388	{ 0.270	Bleaney (1950)
		{ 0.22	{ 0.15 }	

In the figure the theoretical specific heat curve calculated from the theory of Hebb & Purcell for a splitting of 0.245°K is shown in figure 1, curve *C*. It will be seen that near and below the maximum the experimental curve departs markedly from the theoretical. This is not surprising in view of the presence of two splittings reported in the preceding paper. It was pointed out there that the assumption that different sets of ions have different splittings also leads to difficulties. The first assumption, that only 15 % of the ions have $\delta = 0.388^\circ\text{K}$ and the other 85 % $\delta = 0.22^\circ\text{K}$, leads to a specific heat curve hardly different from that in the figure for a single splitting of 0.245°K , as well as being at variance with the intensity ratios in the paramagnetic resonance spectrum. The other assumption, that 30 % have 0.388°K , 30 % have 0.22°K and the remainder a small splitting of 0.05°K gives curve *D* in figure 1. It is clear that any distribution of splittings between different percentages of ions which can be arranged to give the spread in the experimental specific heat anomaly will give too low a value near the maximum. This can be seen more clearly from a consideration of the entropy curve, where de Klerk *et al.* find a plateau, not at $S/R = \log 2$, but about half this value. This requires that some of the entropy corresponding to the Kramers' degeneracy is also removed in passing down through the anomaly. The suggestion of a hyperfine splitting in the 9 % abundant ^{53}Cr isotope has been considered in the preceding paper, where it was pointed out that the paramagnetic resonance spectrum could not be explained by such an effect. Here it may be pointed out that the hyperfine structure would have to be of the same order as the crystalline electronic splitting to cause a material shift in the position of the entropy plateau. Although the observed *hfs* in vanadium is of this order, it seems unlikely to be so large in chromium, for ^{51}V has a very large magnetic moment,† and

† Paramagnetic resonance measurements by Ingram, Scovil and the author show conclusively that the nuclear spin of ^{51}V , hitherto slightly doubtful, is $\frac{7}{2}$. The magnetic moment is therefore 5.15 nuclear magnetons, as calculated by Knight & Cohen (1949) on this assumption.

that of ^{53}Cr , which has an even atomic number, is probably much smaller. Moreover the nuclear spin of ^{53}Cr would have to be enormous, owing to its low abundance, in order to account for the shift in the entropy plateau of the salt as a whole. For these reasons it seems extremely improbable that the shift is due to a hyperfine structure.

6. CONCLUSION

Potassium chromic alum has long been regarded as one of the simpler and more desirable substances for use in experiments below 1°K . Exchange interaction between the paramagnetic ions was expected to be small, and the presumption of only a single crystalline splitting seemed to make the theoretical interpretation of the results relatively simple. In this paper the author has attempted to show that the information obtained from adiabatic demagnetization and paramagnetic resonance investigations is consistent with none of the simple explanations advanced, nor is it easy to reconcile one set of results with the other. It is a great pity that this alum and the chromic ammonium alum appear to be the only substances so far which cannot be investigated by paramagnetic resonance when highly diluted with the corresponding isomorphous diamagnetic substance (aluminium alum). Otherwise, one may confidently anticipate that a great deal more detailed information would have been obtained which might have afforded a clue to this problem.

For work below 1°K , the paramagnetic resonance measurements suggest that the chromic alums of rubidium, caesium and methylamine would offer the best comparison with the theoretical work of Hebb & Purcell, since they appear to have only one splitting. The methylamine alum which is inexpensive and easy to grow in the form of large crystals, seems to be the most suitable. Preliminary experiments have shown that very narrow lines may be obtained in the paramagnetic resonance spectrum of this salt, diluted with the corresponding aluminium alum, and it is hoped to make a thorough investigation and obtain an accurate value for the splitting. If satisfactory agreement can be obtained with adiabatic demagnetization experiments on this salt, it should replace the potassium chrome alum as a 'standard substance' for work below 1°K .

The author wishes to thank Professor F. Simon for suggesting this problem; and Drs N. Kurti and A. H. Cooke, and the late Dr R. A. Hull, for much valuable discussion and help with the experimental work.

REFERENCES

- Bleaney, B. 1950 *Proc. Roy. Soc. A*, **204**, 203.
 Bleaney, B. & Hull, R. A. 1941 *Proc. Roy. Soc. A*, **178**, 74.
 Bleaney, B. & Simon, F. 1939 *Trans. Faraday Soc.* **35**, 1205.
 Broer, L. J. F. 1947 *Physica*, **13**, 352.
 Casimir, H. B. G., Bijl, D. & du Pré, F. K. 1941 *Physica*, **8**, 449.
 Casimir, H. B. G., de Haas, W. J. & de Klerk, D. 1939 *Physica*, **6**, 365.
 Casimir, H. B. G., de Klerk, D. & Polder, D. 1940 *Physica*, **7**, 737.
 Cooke, A. H. 1949 *Proc. Phys. Soc. A*, **62**, 269.
 Gorter, C. J., Dijkstra, L. J. & van Paemel, O. 1942 *Physica*, **9**, 673.

- Hebb, M. H. & Purcell, E. M. 1937 *J. Chem. Phys.* **5**, 338.
Hull, R. A. 1947 *Report of Cambridge Conference on Low Temperatures*, p. 72. London: Physical Society.
de Klerk, D., Steenland, M. J. & Gorter, C. J. 1949 *Physica*, **15**, 649.
Knight, W. D. & Cohen, V. W. 1949 *Phys. Rev.* **76**, 1421.
Kurti, N., Lainé, P., Rollin, B. V. & Simon, F. 1936 *C.R. Acad. Sci., Paris*, **202**, 1576.
Schmidt, G. & Keesom, W. H. 1937 *Physica*, **4**, 971.
Starr, C. 1941 *Phys. Rev.* **60**, 241.
Van Vleck, J. H. 1940 *Phys. Rev.* **57**, 426.
-

The scattering of gamma-rays in extended media

I. Perpendicular incidence on a plane slab

BY L. CAVE, J. CORNER AND R. H. A. LISTON

Ministry of Supply

(Communicated by W. G. Penney, F.R.S.—Received 22 June 1950)

As the simplest problem involving the multiple scattering of gamma-rays, the case of perpendicular incidence on a plane slab has been studied by various simple theoretical methods. No exact solution is available as a test of these methods, but approximations of increasing accuracy, known to give too small a dose under the slab, have been combined with a theoretical upper limit for the dose, forming a zone in which the exact solution must lie. Most of the methods tested could be used even if pair-creation or the photoelectric effect gave an appreciable absorption, though for simplicity the computations have been carried out for pure Compton scattering and for quantum energies of a few MeV. Polarization by the Compton effect has been neglected. The best of the simple methods is a modification of the upper limit solution; in problems where this is too difficult to formulate, a good alternative is to calculate the dose from unscattered and once-scattered quanta, giving the latter the penetrating power of the incident quanta and the maximum possible energy-absorption cross-section.

1. INTRODUCTION

When the scattering coefficient of a slab of material for radiation, such as visible light or gamma-rays, is measured experimentally, it is arranged, if possible, that the scattered radiation shall not reach the detector at the far side of the slab. Such an arrangement is said to have 'good geometry'. Arrangements where scattered radiation has still an appreciable chance of reaching the detector are said to have 'poor geometry'. It is much more difficult to deduce the scattering coefficient from experiments of the second type, and, conversely, if we know the scattering coefficient it is difficult to estimate how much radiation reaches the detector in conditions of 'poor geometry'. In many fields, however, the attenuation of radiation occurs under such conditions; for example, in high-voltage radiography these problems occur frequently, and are there usually called 'wide-beam' conditions. It was the increasing use of gamma-rays in radiography and radiotherapy which led to the investigations reported here. The aim was to find simple mathematical methods for solving such problems and to obtain some idea of the accuracy inherent in the methods, and then to analyze some simple lay-outs which resemble those occurring in radiography.

The penetration of radiation through scattering layers was first studied by Schuster (1905). It was assumed that in the act of scattering the radiation had equal probability of being scattered into any direction ('isotropic scattering'). The only other scattering law which has been used at all widely is the Rayleigh law, which makes the probability (per unit solid angle) proportional to $1 + \cos^2 \theta$, where θ is the angle of deflexion of the radiation. This law is important in accounting quantitatively for the illumination of the sky by scattered sunlight. Exactly the same form occurs in Thomson scattering of light by free electrons, a process which is believed to be important in certain types of hot star. Astrophysics has not up to now needed any highly anisotropic scattering law, nor has the recent development of neutron physics, most of which assumes isotropic scattering. Qualitatively the effect of anisotropy has been investigated by adding Legendre polynomials in $\cos \theta$ to the scattering law. Chandrasekhar (1946*b*) has indeed found an exact solution for the radiation at any point in a semi-infinite slab, illuminated from any direction, when the scattering law can be expressed as a series of Legendre polynomials in $\cos \theta$.

The scattering of gamma-rays occurs in Compton collisions with electrons, for which the scattering law (Klein & Nishina 1929) is highly anisotropic; furthermore, it is not expressible as any short series of Legendre polynomials of *low* order. Chandrasekhar's solution is therefore too laborious to be applied to gamma-ray scattering; in any case, his work assumes no change of wave-length in scattering, whereas in Compton scattering there can be a relatively large change of wave-length, depending on the angle of scattering. Problems of the penetration of gamma-rays have therefore a mathematical as well as a practical interest, being examples of highly anisotropic scattering with large but variable degradation of wave-length.

The problems of practical interest are unfortunately nearly all complicated by the presence of two media and by the geometrical conditions. Thus the mathematical methods must be easy to set up and also easy to compute. In the present paper we compare various methods in these two respects. There is, so far as we know, no exact solution for any gamma-ray penetration problem. As a control, we have taken what we think is probably the easiest problem, that of perpendicular incidence on a slab of material, and have found upper and lower limits to the solution. These limits are sufficiently close to show the relative accuracy obtained by the various approximate methods. This test problem is also not without some practical interest. For simplicity we have omitted absorption by pair-creation and the photoelectric effect, though most of the methods could take these effects into account; it is not expected that the relative merits of the various methods of solution would be altered by including these effects.

There is little previous work to be noted. Hirschfelder, Magee & Hull (1948) and Hirschfelder & Adams (1948) have published an approximate method for the case of perpendicular incidence on a plane slab. Their method could probably be modified to cover other geometrical arrangements. Our results show that this method overestimates the radiation received beyond the slab.

Bobrowsky (1948) has suggested that a thick layer of scattering and absorbing material should be regarded as a pile of layers, each sufficiently thin for single-scattering to be the only important effect. The solution for the thick slab can then

be obtained by repeated matrix multiplication of the elementary solutions for the thin layers. The illustrative problems solved by Bobrowsky do not include any in which there is highly anisotropic scattering, for which '(electronic computing) machine methods for matrix multiplication are certainly necessary'. It is unfortunate that repeated matrix multiplication makes heavy demands on machine capacity. It is probable that anyone with an electronic computer would solve gamma-ray problems by computing from other mathematical solutions.

Bethe, Fano & Karr (1949) have obtained asymptotic formulae for the penetration of gamma-rays through a plane slab. They have treated the radiation, whose wavelength has been little altered, by an expansion in powers of the relative frequency change, neglecting the associated deflexions. Our experience with expansions in terms of the deflexion suggests that this effect is not negligible.

It should be added that the theoretical mechanisms of scattering and absorption of gamma-rays are known to account very well for experiments with good geometry (cf. Cowan 1948). The theoretical difficulties with poor geometry are purely the mathematical difficulties arising in repeated scatterings; the mechanisms of the individual absorptions and scatterings are not in doubt.

Experiments have been carried out by Wyckoff, Kennedy & Bradford (1948), using X-rays up to 1.4 MeV peak applied voltage, and Singer, Braestrup & Wyckoff (1946) have worked at 2 MeV peak, in lay-outs simulating typical protective barriers near high-voltage X-ray machines. Kaye, Binks & Bell (1936) have used lay-outs with poor geometry with a radium source. Such experiments are of direct practical application, though only to the particular geometrical arrangements used, but they do not provide a convenient test of mathematical methods, which are made much more laborious when the incident rays have a number of energies (as for radium) or form a continuous spectrum, as from a high-voltage X-ray tube.

In §2 we summarize the relevant properties of gamma-rays, so that in §3, where we state our standard problem, it will be clear in what ranges of the variables we must work. In §4 we give lower limits for the dose, with an upper limit in §5. Various approximate methods are sketched in §6 and the results compared in §7. Appendix A contains tables of various functions arising from the Klein-Nishina formulae. Mathematical details of some of the methods sketched in the main part of the report are given in appendices B to E.

Throughout we shall assume that the gamma-rays have energies of a few MeV.

2. RELEVANT PROPERTIES OF GAMMA-RAYS

2.1. *Mechanisms of scattering and absorption of gamma-rays*

There are three main mechanisms of absorption and scattering of gamma-rays and X-rays. In the first, the internal photoelectric effect, the photon disappears and an electron is ejected from an atom. The process is a superposition of the effects due to individual electronic levels. The effect occurs only if the energy, $h\nu$, of an incident quantum, is greater than or equal to the binding energy of the electronic level considered. The effect is biggest when $h\nu$ is equal to the binding energy and falls off rapidly at higher $h\nu$, according to the well-known Kramer formula. For elements

of low atomic number the effect is unimportant at all gamma-ray energies. With increasing atomic number the effect becomes more noticeable, thus for iron photoelectric absorption is important only up to $\frac{1}{2}$ MeV whereas for lead it is significant even up to 5 MeV. We neglect photoelectric absorption in the remainder of this paper.

The second mechanism of scattering, and by far the most important for our purpose, is the Compton effect. In this process the reactants are a photon and a free electron at rest.* The process gives a deflected photon of lower energy, settled by the angle of deflexion, and an electron with the surplus energy. The changes of energy are substantial if the deflexion is at all marked. Actually there is a high probability that the deflexion will be small, as with all processes involving electrons at relativistic energies. Figure 1 shows the probability of a deflexion, measured by the radius vector in that direction, when the incident quantum has an energy of about 2 MeV. The directions are marked in which the energy of the scattered quantum has certain simple values.

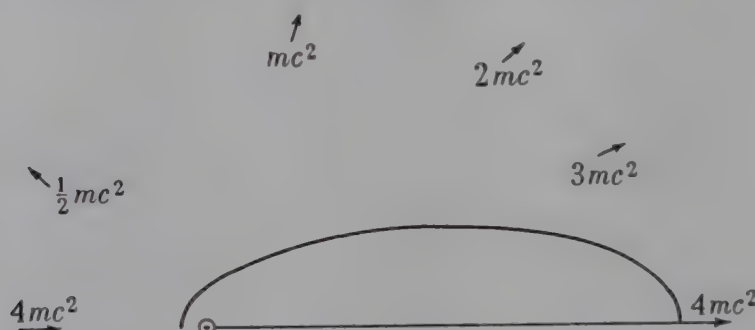


FIGURE 1. Polar diagram for the relative probability of scattering of a $4mc^2$ (~ 2 MeV) quantum, by Compton collision with a free electron. The directions are indicated in which the scattered quanta have certain simple energies.

The anisotropy of the scattering decreases if the incident energy becomes less, and for vanishingly small energy the process passes over to Thomson scattering of light on free electrons. The scattering coefficient then depends on the angle of deflexion θ in the same way as in Rayleigh scattering, being proportional to $1 + \cos^2 \theta$. In this limit the probability of scattering directly backwards ($\theta = -\pi$) is the same as for scattering directly forward ($\theta = 0$). This limiting state is approached very slowly as the quantum energy is reduced; for example, not until this energy falls to 70 keV does the back-scattering probability rise to 50 % of the forward-scattering probability.

A third mechanism of absorption, that of pair-creation, comes into play when the photon energy is greater than 1.022 MeV. In pair-creation a photon passing a nucleus disappears, leaving a pair of electrons, positive and negative, carrying the surplus energy above the $2mc^2$ needed to create the pair. Here m is the electron mass and c the velocity of light; mc^2 , the natural unit of energy when considering Compton and pair-creation effects, is 0.511 MeV. The pair-creation effect begins, then, at $h\nu = 1.022$ MeV, with zero probability at this energy. The behaviour is discussed by,

* When the photon is a gamma-ray the binding energy of even a K -electron is of order $10^{-6}Z^2$ of the energy of the photon. The ratio is less than 1 % even for lead, and therefore all electrons can be considered as free and at rest.

for example, Heitler (1944), where it is shown that the probability, for a nucleus of atomic number Z , is proportional to Z^2 . The Compton effect per atom is proportional to Z , so that the ratio of pair-creation to Compton effects is itself proportional to Z . This ratio also increases rapidly with the photon energy. At the energies of interest in radiography, say 2.5 MeV, the ratio of pair-creation to Compton cross-section is only 2 % for air and 4 % for concrete, though it rises to 25 % for lead.* We shall therefore omit pair-creation from our test example. The effect can, indeed, be included approximately in most of the mathematical methods used, but would involve rather more computation. It is expected that the relative merits of the various methods would be the same if applied to a substance, such as lead, in which pair-creation is important.

Other absorption and scattering mechanisms are mostly high-order processes, and the theoretical conclusion that they would be trivial in practice has been verified by Cowan (1948).

2.2. Formulae for the Compton effect

We define first the cross-section for scattering into a given direction. Let a single electron be subject to a beam of quanta, of number I per unit area. Then the number of quanta scattered into a small solid angle $d\Omega$ with deflexion θ is written as $Ig(\mu, \gamma) d\Omega$, where μ is an abbreviation for $\cos \theta$ and γ is $h\nu/mc^2$ for the incident quanta. The Klein-Nishina formula is

$$g(\mu, \gamma) = \frac{1}{2}r_0^2 \left[\frac{1 + \mu^2}{\{1 + \gamma(1 - \mu)\}^2} + \frac{\gamma^2(1 - \mu)^2}{\{1 + \gamma(1 - \mu)\}^3} \right], \quad (1)$$

where r_0 is the classical 'electron radius' $e^2/mc^2 = 2.80 \times 10^{-13}$ cm. At $\mu = 1$, $g(\mu, \gamma)$ is always r_0^2 . For gamma-rays of a few MeV, $g(\mu, \gamma)$ drops rapidly as μ decreases from 1, that is, the probability of scattering decreases rapidly as the angle of deflexion increases, and this drop is the more rapid the larger γ . Figure 1 shows $g(\mu, \gamma)$ for $\gamma = 4$, corresponding to quanta of about 2 MeV. For the limit of low-energy quanta, $\gamma = 0$, the scattering law degenerates into that for Thomson scattering of light by free electrons:

$$g(\mu, \gamma) = \frac{1}{2}r_0^2(1 + \mu^2).$$

The total number of quanta scattered in all directions is $I\sigma(\gamma)$, where $\sigma(\gamma)$ is the Compton total cross-section for quanta with $h\nu/mc^2 = \gamma$. By integration of (1)

$$\sigma(\gamma) = \pi r_0^2 \{ (\gamma - 2 - 2/\gamma) \ln(1 + 2\gamma) + 4 + \gamma/2 - \gamma/2(1 + 2\gamma)^2 \} / \gamma^2. \quad (2)$$

This is tabulated in appendix A for γ from 0.2 to 10. In this range σ decreases steadily from $6.2r_0^2$ to $1.0r_0^2$, that is, from 4.8 to 0.8×10^{-25} cm.².

Quanta scattered through a deflexion θ , with $\mu = \cos \theta$, emerge with an energy given by $h\nu/mc^2 = E$, where

$$E = \gamma/(1 + \gamma(1 - \mu)). \quad (3)$$

Thus back-scattered quanta have $E \simeq \frac{1}{2}$, corresponding to about $\frac{1}{4}$ MeV, unless γ is small compared to unity.

* The cross-section usually quoted for pair-creation assumes that the atom is left in its original state. It is possible also to leave the atom in an excited state. Wheeler & Lamb (1939) have shown that for cosmic-ray energies and air the effect increases the pair cross-section by 17 %. The effect is smaller in our range of energies and is certainly not sufficient to make pair-creation an important factor.

One is often concerned more with the flux of energy rather than numbers of quanta. It is then convenient to use a differential cross-section for the scattering of energy, defined analogously to $g(\mu, \gamma)$: let an electron be subject to a beam of energy H per unit area; then the energy scattered into a small solid angle $d\Omega$ with deflexion θ is written as $Hf(\mu, \gamma)d\Omega$ and, from (3),

$$f(\mu, \gamma) = g(\mu, \gamma)/\{1 + \gamma(1 - \mu)\}, \quad (4)$$

so that with (1)
$$f(\mu, \gamma) = \frac{1}{2}r_0^2 \left[\frac{1 + \mu^2}{\{1 + \gamma(1 - \mu)\}^3} + \frac{\gamma^2(1 - \mu)^2}{\{1 + \gamma(1 - \mu)\}^4} \right]. \quad (5)$$

The notation $f(\mu)$ has been used by Hirschfelder *et al.* (1948), who have tabulated $f(\mu, \gamma)$ for $\gamma = 0.5, 1, 2, 6$ and 10 (table II of their paper). They tabulate also the useful function $\sigma(E)$, where E arises from γ and μ by (3), and $\sigma(E)$ is the function (2) with γ replaced by E . In their notation $\sigma(E)$ is σ_1 . It is the Compton cross-section for quanta which have already been scattered once through a deflexion θ , with $\mu = \cos \theta$.

An incident quantum with energy γmc^2 is changed by scattering into a quantum of energy $E mc^2$. The energy $(\gamma - E) mc^2$ appears as kinetic energy of the recoil electron and eventually as ion pairs and excited atoms. Thus from an incident beam of energy H per unit area the amount of energy lost from the radiation field by scattering into $d\Omega$ is

$$H(1 - E/\gamma)g(\mu, \gamma)d\Omega = \gamma H(1 - \mu)f(\mu, \gamma)d\Omega, \quad (6)$$

and the total amount of energy lost, due to the presence of a single electron, is $H\sigma_a(\gamma)$, where σ_a is found, by integration of (6), to be

$$\sigma_a = \pi r_0^2 [(\gamma^{-1} - 2\gamma^{-2} - 3\gamma^{-3}) \ln(1 + 2\gamma) - 8\gamma^2/3(1 + 2\gamma)^3 + (6 + 22\gamma + 18\gamma^2 - 2\gamma^3)/\gamma^2(1 + 2\gamma)^2]. \quad (7)$$

σ_a is the 'cross-section for energy absorption by Compton scattering', and is tabulated in appendix A for γ from 0.4 to 10 . This cross-section has a maximum of $1.2431r_0^2$ at γ very close to unity, corresponding to 0.5 MeV quanta. σ_a is needed as a link between the energy in a stream of radiation and the effects produced in a body of low atomic number.

Formulae such as (1) assume that the incident radiation is unpolarized, since they are obtained by averaging the original Klein-Nishina formula which took polarization into account. Let θ be the angle of scattering and ψ the angle between the directions of polarization of the incident and scattered radiation. Then the number of quanta scattered into $d\Omega$ with this polarization is really

$$I r_0^2 (E/\gamma)^2 [(E/\gamma) + (\gamma/E) - 2 + 4 \cos^2 \psi] d\Omega/4. \quad (8)$$

Let $\mathbf{k}_0$ (figure 2) be a vector in the direction of motion of the incident quanta, $\mathbf{k}$ in the direction of the scattered quanta. Let $\mathbf{e}_0$ be the direction of polarization of the incident quanta, and let ϕ be the azimuthal angle of $\mathbf{k}$ relative to the plane $(\mathbf{k}_0, \mathbf{e}_0)$. Take standard directions of polarization $\mathbf{E}_1, \mathbf{E}_2$ respectively in and perpendicular to the plane $(\mathbf{k}_0, \mathbf{k})$. Then

$$\begin{aligned} \cos^2 \psi &= \cos^2 \phi (1 - \sin^2 \theta \cos^2 \phi) \quad \text{for } \mathbf{E}_1 \\ &= \sin^2 \phi (1 - \sin^2 \theta \cos^2 \phi) \quad \text{for } \mathbf{E}_2. \end{aligned}$$

Now we allow the incident light to be unpolarized. The vector $\mathbf{e}_0$ takes all possible directions perpendicular to $\mathbf{k}_0$, while the directions $\mathbf{E}_1$ and $\mathbf{E}_2$, defined as above, remain fixed. This amounts to a simple averaging of the intensities over ϕ . The ratio of the average intensities in directions $\mathbf{E}_1$ and $\mathbf{E}_2$ is

$$\frac{(E/\gamma) + (\gamma/E) - (\frac{3}{2})\sin^2\theta}{(E/\gamma) + (\gamma/E) - \frac{1}{2}\sin^2\theta} = 1 - \tau, \text{ say.}$$

We have $\tau \simeq \frac{1}{2}\theta^2$ for small deflexions. Most of the scattering takes place for smaller deflexions than are given by

$$\mu = 1 - 1/\gamma,$$

for at this angle $g(\mu)$ is already less than a third of $g(\mu)$ at $\mu = 1$. At $\mu = 1 - 1/\gamma$

$$\tau = \frac{2\sin^2\theta}{5 - \sin^2\theta} \simeq 4/(5\gamma - 2).$$

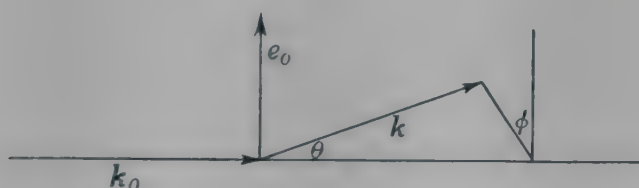


FIGURE 2. Notation for directions of motion and polarization in Compton scattering.

At $\gamma = 4$, $\tau \simeq \frac{2}{9}$. The polarization produced by single Compton scattering of an initially unpolarized beam is therefore small compared to the marked anisotropy of direction and wide spread of resulting quantum energies. This is a justification for neglecting polarization in this paper while keeping the anisotropic scattering coefficient and the degradation of energy.*

2.3. Radiation measurements

Ideally, one would like to know the radiation at any point as a distribution over angle and wave-length. This would involve much computation, and it seems more reasonable to choose some single parameter to represent the radiation at any assigned point. The obvious parameter is the energy released per unit volume at the point, since this energy can be measured with a non-directional counter.

A Röntgen of gamma- or X-radiation is defined as that amount which liberates in 1 cm.³ of air at N.T.P. ions carrying 1 electrostatic unit of charge of either sign. This amounts to a liberation of 2.08×10^9 ion-pairs per cm.³ requiring

$$33.2 \times 2.08 \times 10^9 \text{ eV/cm.}^3 = 6.9 \times 10^4 \text{ MeV/cm.}^3.$$

This energy is provided by Compton scattering and pair-creation. The energy given to the recoil electron in the former process is measured by σ_a , the energy absorption

* The only problem involving polarization effects which has ever been solved is the case of Rayleigh scattering, first solved exactly by Chandrasekhar (1946a). Polarization introduces very considerable difficulties into an exact analytical solution and would probably do the same for our analytical approximations and computations. However, it seems clear that before one could hope to get a solution closer to the truth than the limits assigned in this report it would be necessary to make an estimate of polarization effects. Wightman (1948) has discussed the polarization to be expected after two Compton scatterings.

cross-section, defined in § 2.2, but some of this energy is radiated as 'bremsstrahlung'. The effect is discussed by Fowler, Lauritsen & Lauritsen (1948), who state that the effect is to add to σ_a a factor $1 - Zh\nu/5000 mc^2$, where Z is the mean atomic number of the medium. For air, water and the human body $Z \sim 7$, while $\gamma = h\nu/mc^2$ is at most 10 in radiographic work. The correction is therefore of order only 1 %.

The effect of a pair-creation is to replace a quantum of energy $h\nu$ by two quanta with total energy $2mc^2$. The cross-section for energy absorption by pair-creation is therefore $(1 - 2mc^2/h\nu) \times$ the cross-section for pair-creation. Also for quanta of a few MeV the relative magnitudes of Compton energy absorption and scattering are roughly $\sigma_a \simeq \frac{1}{2}\sigma$. The net effect is to make energy absorption by pair-creation almost exactly as trivial compared with the Compton effect as we have already seen to hold for loss of quanta out of a beam.

Consider now a small volume $d\tau$, in which we require the energy absorption from a radiation field which depends on direction. It is necessary first to define radiation 'intensity', which can be related either to energy passing or number of quanta. There is, of course, no physical significance in the choice. In the present context it is more convenient to define 'intensity' as a measure of the number of quanta passing. Take, then, a small dA at a point P , with normal in direction $\mathbf{n}$, and let the number of quanta crossing this area per unit time with directions of motion in a small solid angle $d\Omega$ around the direction $\mathbf{n}$ be written as $I(\mathbf{n}, P) dA d\Omega$. This defines the intensity I , which in general depends on two angles and three space co-ordinates. We define directions by spherical polar co-ordinates (θ, ϕ) , or (μ, ϕ) where $\mu = \cos \theta$. Then if we write $E = h\nu/mc^2$, the energy received by the small volume $d\tau$ from radiation coming from a solid angle $d\Omega$ in direction (μ, ϕ) is, by definition of σ_a ,

$$Emc^2 I(\mu, \phi) N \sigma_a(E) d\tau d\Omega,$$

where N is the number of electrons, free or bound, per unit volume. The energy per electron from all directions is

$$E \sigma_a(E) mc^2 \int I(\mu, \phi) d\Omega,$$

while if the radiation covers a continuous spectrum of values of E we define I per unit range of E and obtain for the energy per electron

$$mc^2 \int E \sigma_a(E) dE \int I(\mu, \phi, E) d\Omega. \quad (9)$$

We shall refer to (9) as the 'dose'. It is related to the constitutional effects produced in human beings by whole-body irradiation, and is therefore important in considering scattered radiation from high-voltage X-ray and gamma-ray sources. The reason is that general constitutional effects depend on the integral of the dose over all the absorbing elements of the body (cf. Mayneord 1940; and Mayneord & Clarkson 1944*a*). The effects are, it is true, a little less if the dose is received in a relatively small part of the body (Ellis 1946), or if the irradiation is spread over a period of weeks or more. If now we calculate the integral dose we find it is proportional to (9) provided that $N \sigma_a L$ is small compared to unity, where L is the average travel of a quantum in the human body. Provided we are not dealing with very unsymmetrical fields, we

can use the experimental result (Mayneord & Clarkson 1944*b*) that for the gamma-rays used in radiotherapy $N\sigma_a L$ is of order one-third.

A difficulty arises from the form of (9) when we consider radiation incident on a slab of finite thickness but of infinite extent. The intensity is constant along a direction not encountering scattering or absorbing material. The intensity occurring in the dose (9) can therefore be taken at the boundary of the slab instead of at the counter position. In any practical application there would be a cut-off for rays too nearly parallel to the screening slab, for these would intercept walls or thick supports before reaching the slab. In such screened directions $I(\mu, \phi)$ at the slab would be finite but at the point of measurement would be effectively zero. This cut-off would depend on the particular circumstances. To get an idealized problem suitable for mathematical investigation we omit this cut-off. As it happens, this leads to some difficulties with limits for very thin slabs, but these are purely mathematical troubles. In practice there is always some cut-off, and in any event the mathematical difficulties begin to appear only for slabs less than 0.1 mean free path thick, which are too thin to be interesting.

Hirschfelder *et al.* (1948) computed not (9) but

$$mc^2 \int E \sigma_a(E) dE \int I(\mu, \phi, E) \mu d\Omega, \quad (10)$$

where θ is measured from the normal to the slab. This reduces the difficulties in the cut-off region near $\mu = 0$, and also helps their series result to converge. (10) is related to the flux of energy across a surface rather than to the amount absorbed in a limited volume.

3. THE STANDARD PROBLEM

As the simplest test case we take a plane slab of uniform thickness and extending to infinity in its own plane. The incident gamma-rays are perpendicular to the plane of the slab, all have $h\nu/mc^2 = \gamma$, and the number per unit area of the slab is the same at all points of the face.

We omit photoelectric absorption, pair-creation, and polarization effects in the radiation scattered by the Compton effect.

There are actually two possible standards with which to compare the dose beyond the slab. We might take the dose in free air, without any slab present, or we might prefer the dose at the front face of an infinitely thick block of the screening material. These differ by an amount of order 1 % for the values of $\gamma = h\nu/mc^2$ which are important in radiography, say $\gamma = 2$ to 5. A large component of low energy means a large back-scattering effect and so a considerable difference in the two standard doses. As such radiation is only weakly penetrating the effect is not important here, and we shall use one or other of the two possible standards as we please, with possible inconsistencies of no more than 1 or 2 %.

Let D be the thickness of the slab in mean free paths, computed at $h\nu/mc^2 = \gamma$. We represent relative doses in the form $e^{-\alpha D}$, where α will, in general, depend on both γ and D . Radiation passing through the slab without a collision gives a relative dose

e^{-D} , so that α is certainly less than unity. α must also be zero or positive. It will appear that α increases monotonically towards unity as D increases from a small value to very large values.

At larger values of D , a small error in α corresponds to larger relative errors in the dose. This tends to compensate for the increasing theoretical difficulties for thick slabs, so that we may expect, in a general way, to attain from any one approximate method much the same accuracy in α at all values of D .

4. LOWER LIMITS FOR THE DOSE

Lower limits can be computed in a systematic way by evaluating first the radiation which passes through the slab without being scattered, then that which suffers one collision, two collisions, and so on. It is possible to write down the operator which transforms the intensity of n times scattered radiation at all depths, angles and values of $h\nu/mc^2$ into the intensity of $(n+1)$ times scattered radiation at any required depth, angle, and value of $h\nu/mc^2$, and this was done first by Ornstein (1937) for isotropic scattering and no degradation of energy. If we need only the relative dose beyond the slab from $(n+1)$ times scattered quanta this can be computed as a multiple integral requiring a fairly complete knowledge of the dependence of the n times scattered radiation on depth, direction, and $h\nu/mc^2$. The difficulties are purely computational, but these have been a serious barrier to the application of the method in the past. Even for simple scattering laws only one published computation going past the twice-scattered component is known to the authors, and this result, due to van der Hulst (1948), goes only to the thrice-scattered intensity and does not pretend to be more than a rough estimate.

Coming now to such calculations for the Compton scattering law, the highest order known to have been reached when the present work was begun, was the first-scattered radiation, computed by Hirschfelder *et al.* (1948) for a series of values of γ and D . This requires integration over only one variable. The second-order relative dose needs the evaluation of triple integrals and the third-order component involves a quintuple integral. Higher orders do not need more than quintuple integrals, but as these involve the computations of lower order and the limits of integration become rather involved the labour is very heavy. The simplifications in the first- and second-order computations arise from our assumption that the incident radiation has a single direction and only one value of $h\nu/mc^2$. After one scattering there is a spread in direction, but at each angle there is still only one value of $h\nu/mc^2$. After the second scattering the spectrum takes a continuous distribution. Third and higher orders are all essentially alike. The breaking up of line spectra in energy and direction is marked mathematically by the disappearance of Dirac δ -functions. The formulae are given in appendix B.

It was thought to be worth computing the twice-scattered dose, which was computed for $\gamma = 10$ and $D = 0.1, 0.2, 1, 4$ and 16 . Once-scattered doses were also computed, for our definition of dose, for a range of values of γ and D . An attempt was made to estimate the three-times scattered dose at $\gamma = 10$ and $D = 16$. An accurate computation would have involved too much work, but by using the experience

gained with the computation of twice-scattered radiation it was possible to get an estimate, thought to be correct to better than 50 %, in a reasonable time. Even a single point obtained in this way is valuable, because the curves of α against D from successive approximations have nearly the same shape. The results of these computations are given in §7.

5. UPPER LIMITS FOR THE DOSE

It is desirable to get at least one upper limit for the relative dose. We shall describe a solution which is an upper limit for slab thicknesses greater than some critical value in the neighbourhood of two mean free paths. Certainly the solution is an upper limit for thicknesses greater than three mean free paths. The method has its peculiarities, it is true; for D , the slab thickness, greater than about 6 (for $\gamma = 10$), the value of α begins to decrease as D increases; this is not plausible, and a better upper limit for the dose (that is, a lower limit for α), is obtained at larger D by using the value of α near $D = 6$. For very small D , of the order of 10^{-5} , the value of α proves to be negative; this indeed gives an upper limit to the dose, but not a very helpful one. As such small values of D are of no importance, the defect is trivial.

The integral equation for this upper limit is obtained by a process which considers single scatterings only. This does not mean that multiple scatterings do not contribute to the dose, but only that they are included implicitly rather than explicitly. It is possible to set up a process which makes explicit reference to both first and second scatterings, and this would give a closer upper limit to the dose. It appears, however, that the region in which the solution fails to be an upper limit would extend to larger thicknesses than the previous value of about three mean free paths. We have not evaluated this more accurate solution, since even the upper limit based on single scatterings is roughly as laborious to compute as the lower limit using twice-scattered radiation.

The method of arriving at an upper limit is given in appendix C. In the present section we give only a sketch of the method. We start with the relative dose $e^{-\alpha D}$ at thickness D mean free paths, where α depends on γ and D . Of this dose e^{-D} comes from radiation which has not been scattered and $e^{-\alpha D} - e^{-D}$ from radiation scattered at least once. We now write down the scattered radiation arriving at a point on the far side of the slab, labelling the quanta according to the point at which they last had a collision. At such a point we know the unscattered component in direction, intensity and wave-length, and we know the scattered radiation only as regards total dose. We assume an average direction and wave-length for this scattered component and proceed to calculate the total scattered radiation at a point on the far side of the slab. It remains to choose the average direction and wave-length, already mentioned, to maximize the dose at the far side of the slab. The simplest method is to assume the original direction and initial $h\nu/mc^2$. It can be shown that this fails to maximize the dose for small D but ought to be suitable at large D . Finally, the scattered dose $e^{-\alpha D} - e^{-D}$ is obtained as an integral involving the function α for the initial value of $h\nu/mc^2$ and all thicknesses up to D . This integral equation has been converted to an integro-differential equation and solved numerically for $\gamma = 10$.

It will probably be clear from this sketch that we are using interchangeably the relative dose at a depth D in a semi-infinite block and at the far side of a slab of thickness D . This is permissible in our work, where errors of 1 % are trivial and we are interested only in fairly penetrating radiation. The error would vitiate the method if the scattering were isotropic or nearly so.

6. APPROXIMATE SOLUTIONS

Various approximate solutions have been tried. The virtues looked for are ease of formulation in complicated geometrical conditions, ease of computation, and accuracy. The most obvious course is to use one of the simpler of the upper or lower limits. The dose from unscattered and once-scattered radiation is easily computed. A simple allowance for high-order terms can be made by taking the scattering cross-section to be the same for the scattered as for the incident radiation, while giving σ_a its maximum possible value $1.243r_0^2$, wherever it occurs in the integral for the once-scattered radiation dose; we use, however, $\sigma_a(\gamma)$ if γ is less than unity. We shall refer to this as the 'modified first-order method'.

It will usually be a virtue if the approximate solution gives too big a dose. This suggests some simplified version of the upper limit discussed in §5. It has been found, indeed, that if we assume that $\alpha(D)$ is independent of D inside the integral mentioned earlier, then the computation becomes quite simple and the solution appears to be relatively accurate. Results will be given in §7. This 'modified upper limit' solution is set up in appendix C.

Hirschfelder *et al.* (1948) have evaluated exactly the doses from unscattered and once-scattered radiation, and have deduced from the latter an 'average scattering angle' for various thicknesses of slab and values of $h\nu/mc^2$. These average deflexions are then used to determine the intensity of the multiple-scattered radiation. The method depends on these deflexions being small and on their slow variation with $h\nu/mc^2$ and slab thickness, and as published applies only to radiation at normal incidence to the slab. Some conversion of their results is needed before they become comparable with ours. Hirschfelder *et al.* based their dose on $\int E dE \int I \mu d\Omega$, whereas we use $\int E \sigma_a(E) dE \int I d\Omega$. We have replaced their exact once-scattered term by our exact integral and converted their approximate second and higher terms* by division by the appropriate mean values of the deflexion μ and multiplication by $\sigma_a(E)/\sigma_a(\gamma)$, where E is the mean $h\nu/mc^2$ for the term under consideration. The only difficulty lies in the slow convergence, which is poor even for thicknesses as small as two mean free paths, and we have not attempted to estimate the contribution from radiation scattered more than 10 times; such radiation we have simply ignored. At $D = 20$ the result so calculated was clearly so arbitrary that we have not plotted it in the graphs of §7. Even the curtailed series gives too big a dose, as will be seen later.

* We are indebted to Professor J. O. Hirschfelder for the private communication of the scattered components up to the tenth order.

The scattering function $g(\mu, \gamma)$ has been introduced in §2. For small $1 - \mu$

$$g(\mu, \gamma) \simeq r_0^2 e^{-2\gamma(1-\mu)}, \quad (11)$$

and this does indicate the rapid decrease of g as μ decreases from unity. For large $1 - \mu$, (11) gives too small a scattering. Similarly for the function $f(\mu, \gamma)$, also introduced in §2, we have the approximation

$$f(\mu, \gamma) \simeq r_0^2 e^{-3\gamma(1-\mu)}. \quad (12)$$

This suggested the following method which we have called the 'exponential approximation without degradation'.

The equation of transfer for the problem is written down, neglecting the degradation of energy in scattering. If this were really true, g and f would be identical; to allow for this error, the scattering law is replaced by $r_0^2 e^{-\beta(1-\mu)}$, with β arbitrary at this stage. The unscattered radiation is split off and an equation of transfer is obtained for the intensity of radiation which has been scattered at least once. A plausible form with two arbitrary functions is assumed for this intensity and inserted in the transfer equation, which is expanded in powers of $1 - \mu$. The terms in $(1 - \mu)^0$ and $1 - \mu$ give two differential equations which can be solved for the arbitrary functions. Finally, the relative dose is obtained by an integral over the scattered intensity and the usual term from the unscattered radiation. The success of the method depends on guessing a good form for the scattered intensity. The form we used was

$$J(x, \mu) = (\beta/2\pi) A(x) e^{-N\sigma x} e^{-\beta(1-\mu) B(x)}$$

for the scattered intensity J at deflexion $\mu (= \cos \theta)$ and depth x into a semi-infinite block; here $A(x)$ and $B(x)$ are the two arbitrary functions. This solution is not applicable at small values of $N\sigma x$, since close to the surface any scattered radiation is bound to be directed outwards. At bigger depths this difficulty disappears, and the form chosen has the correct tendency to favour radiation going on with small deflexions. The details of the solution are given in appendix D.

An attempt was made to set up an 'exponential approximation allowing for degradation of energy'. The transfer equation was set up and now, to get rid of Dirac δ -functions, not only the unscattered but also the once-scattered radiation was split off. The remaining intensity is now a function of three variables, namely, direction, depth and wave-length, which does not make a suitable form at all obvious. Nevertheless, a plausible form was found and the arbitrary functions obtained by solution of the transfer equation for small deflexions and small-energy degradations. When the relative dose was evaluated it became clear that a major part came from other regions of the variables, and the final result was rather arbitrary. At this stage of the work the computed intensities of twice-scattered quanta became available, and these confirmed that there was probably no simple formula approximating to the intensity of radiation scattered more than once. The method was therefore left in the form given in appendix E.

7. COMPARISON OF SOLUTIONS

Computation began at $\gamma \equiv h\nu/mc^2 = 10$, corresponding to about 5 MeV. Some of the methods are clearly best at large γ , and it was expected that these might be eliminated after seeing their best possible performance.

The upper and lower limit results are shown in figure 3. An upper limit to the dose means of course a lower limit for α . The result for unscattered radiation alone is $\alpha = 1$, and at small values of the slab thickness, D mean free paths, the lower limit results keeping radiation scattered zero, one and two times form a sequence from which it is possible to estimate the result keeping all orders of scattering. At D of order 10 this estimation is not easy, and the result at $D = 16$ to the next order of scattering is particularly useful. From this result and the trend of the curves of lower order an estimated curve for scatterings up to three times has been sketched in.

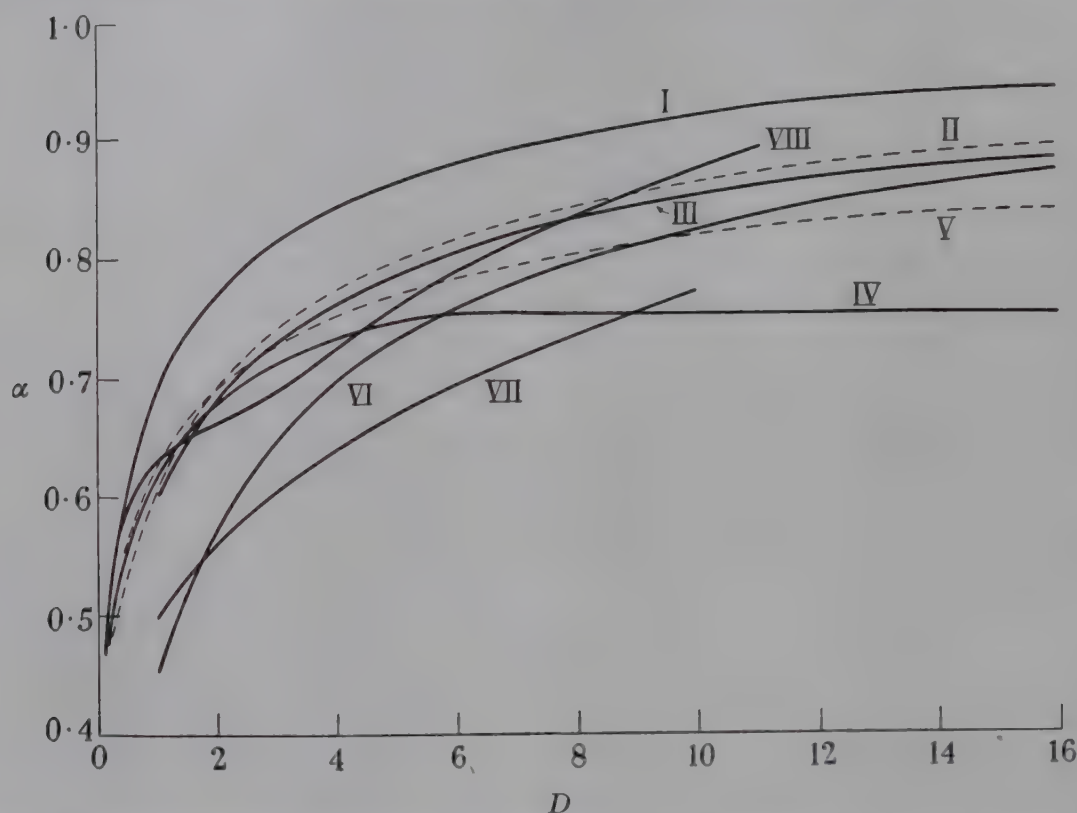


FIGURE 3. Values of α from various methods, for $h\nu/mc^2 = 10$. The relative dose beyond a slab of thickness D mean free paths is $e^{-\alpha D}$. Curve I is for unscattered and once-scattered radiation, curve II is for radiation scattered up to twice, curve III for radiation scattered up to three times. Curve IV is the 'upper limit' (lower limit to α) and curve V is the 'modified upper limit' of §5 and appendix C. Curve VI is the 'modified first-order' approximation and VII is the Hirschfelder result, both described in §6. Curve VIII is the 'exponential approximation with degradation' (appendix E) with $\nu = 0.15$ and $\sigma_T/\sigma = 1.459$.

Curve IV is the 'upper-limit' theory of §5 and appendix C. It is obviously not a true bound until D reaches about 3, and for D greater than 6.5 the theory indicates only that α must be bigger than 0.753. This leaves rather a wide gap between the upper and lower bounds to α . The 'modified upper limit' lies close to the 'strict upper limit' curve for small D (curves IV and V of figure 3), and it appears that the true α cannot lie far below curve V at large D . Indeed, this 'modified upper limit' solution seems to give remarkably good results for very little computation. The only drawback foreseen is that it might be difficult to formulate the equation in more complicated geometrical conditions.

We have mentioned that the dose from radiation scattered zero and one times can be considerably improved and the computation simplified by assuming that (i) the

Compton cross-section is the same for the scattered radiation as for the incident radiation, and (ii) that the energy-absorption coefficient $\sigma_a(E)$ which occurs in the integral for the once-scattered dose, with $h\nu/mc^2 = E$ a function of γ and angle of deflexion, can be replaced by the maximum possible value $\sigma_a = 1.243r_0^2$, if γ is greater than unity, and otherwise by $\sigma_a(\gamma)$. The improvement due to (i) is of the same order at all γ ; from (ii) the change depends a good deal on γ , being biggest at large γ and possibly even overestimating the true dose. The result for $\gamma = 10$ is plotted as curve VI of figure 3. This overestimates the dose (underestimates α) for D less than 8 and gives rather too small a dose at the biggest D , say 16, of interest to us. This approximation is not only easier to compute than the accurate integral for zero- and once-scattered radiation, but is actually more accurate for slab thicknesses greater than two mean free paths.

Curve VII of figure 3 shows the results of Hirschfelder *et al.* converted as described in §6. They give clearly too small a value of α , that is, too big a dose. There is some indication that this might not be the case at really big D , say $D = 20$, where, however, convergence difficulties prevent our quoting a definite result. Computation from this method is relatively heavy, and in the published method there is an approximation which seems to be valid only in the problem discussed in the present report. It should be remembered that Hirschfelder *et al.* were interested in the flux $\int I\mu d\Omega$ and not in $\int Id\Omega$, and that their method would give more accurate results for the former quantity.

The two 'exponential approximations' of §6 gave disappointing results. In the 'exponential approximation without degradation' there are two arbitrary parameters σ , the total cross-section per electron, and β , which must be of order 2γ . It is expected that σ will not differ greatly from the Compton cross-section for scattering at $h\nu/mc^2 = \gamma$. It is found that for most values of σ and β this theory gives an α which decreases as D increases. This is certainly the wrong behaviour, and it does not seem to be possible to eliminate it without getting in exchange a big error at all thicknesses. Indeed, curves of rising α against D appear to lie entirely in the region $\alpha < 0.6$ and so correspond to gross overestimates of the dose. The method has the other disadvantages that it is not particularly easy to set up in complicated geometries and the computation promises to be quite long in such cases.

The 'exponential approximation with degradation of energy' promises to be rather better because at least the contributions of unscattered and once-scattered radiation are obtained exactly. In this method there are two arbitrary parameters: σ_r , the total cross-section per electron, and ν , a parameter arising from the approximation to the scattering law:

$$g(\mu, E) \simeq r_0^2 e^{-2\nu E(1-\mu)}. \quad (13)$$

We expect ν to be of order unity. The approximation (13) gives too small a scattering at large deflexions, and therefore ν should be less than unity; an argument can be given which is based on the total amount of energy scattered and which suggests that ν might be as small as 0.15. There is actually not a lot to be gained by any other choice of ν , since this alters only one term in the total dose. We have therefore used

$\nu = 0.15$ and adjusted σ_τ/σ for best fit to the other theories. At $\sigma_\tau/\sigma = 1$, α decreases as D increases from 1 to 5, and while α does increase thereafter it remains much too small, around 0.6. By going to σ_τ/σ about 1.5 the dip in the $\alpha:D$ curve is almost eliminated and the result (curve VIII of figure 3) is not unsatisfactory. A smaller value of σ_τ/σ would give better agreement at large D but would introduce a dip in the neighbourhood of $D = 2$.

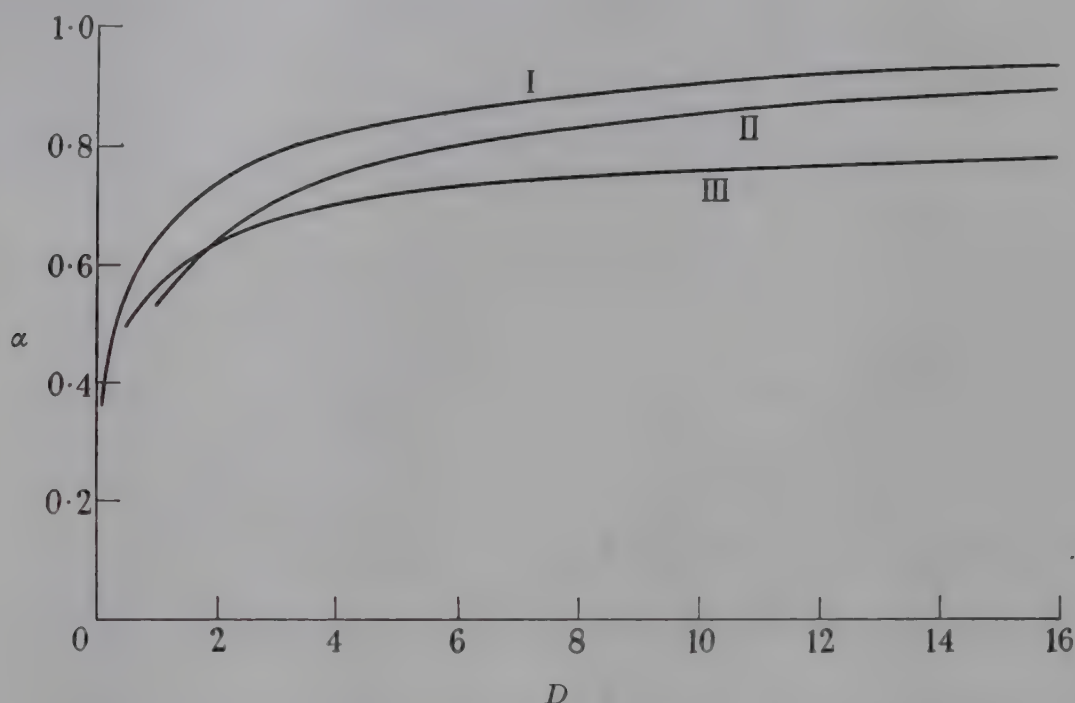


FIGURE 4. Values of α from various methods, for $h\nu/mc^2 = 4$. The relative dose beyond a slab of thickness D mean free paths is $e^{-\alpha D}$. Curve I refers to unscattered and once-scattered radiation. Curve II is the 'modified first-order' theory, while curve III is the 'modified upper limit' result.

All these results refer to $\gamma \equiv h\nu/mc^2 = 10$. On their evidence we may omit the two 'exponential approximation' theories from further consideration, for the only one which is partly successful gives a poor return for the labour involved. The 'upper limit' of §5 and appendix C is laborious to compute and not very useful at large or small D . The 'modified upper limit' is not difficult to compute and, for $\gamma = 10$, lies close to the true solution. The Hirschfelder solution gives too big a dose and the computation seems to be fairly heavy. The lower limits to the dose are useful in fixing the position of the true solution but the labour in finding the dose from radiation scattered two and three times is very heavy.

Some results from the simpler methods, for $\gamma = 4$, are shown in figure 4. The dose from radiation scattered zero and one times gives an α which at large D differs from that for $\gamma = 10$ and the same method by an amount of order 0.01. For smaller D the difference increases, to about 0.035 at $D = 2$. The improvement due to the 'modified first-order' approximation is a good deal smaller at $h\nu/mc^2 = 4$ than at $h\nu/mc^2 = 10$. This arises from $\sigma_a(4)$ being nearer to the maximum value of σ_a than $\sigma_a(10)$. The considerable improvement obtained by the modified first-order method at $\gamma = 10$ is therefore rather an accident; nevertheless, the improvement always exists, and the computation is extremely simple in this approximation.

The 'modified upper limit' gives for $\gamma = 4$ an α which lies below that computed by the same method for $\gamma = 10$. No difficulties of computation arise at the smaller values of γ .

From these results it is possible to choose a method suitable for more complicated situations, with at least some idea of the error involved at various $h\nu/mc^2$ and average path length.

We are indebted to Dr W. G. Penney, F.R.S., for his encouragement of this work, and to the Chief Scientist, Ministry of Supply, for permission to publish.

APPENDIX A. FUNCTIONS ARISING IN COMPTON SCATTERING

The tables in this appendix are designed for interpolation by the formula

$$f(a + nh) = nf(a + h) + (1 - n)f(a) + B''(n)(M''_0 + M''_1) + B'''(n)\Delta''', \quad (14)$$

where $B''(n)$ and $B'''(n)$ are the second and third Bessel interpolation coefficients, tabulated in, for example, *Interpolation and allied tables* (H.M.S.O. 1936). $M''_0 + M''_1$ and Δ''' are tabulated with $f(a)$. Δ''' is the third difference of $f(a)$ and $M''_0 + M''_1$ is the

TABLE A 1. COMPTON TOTAL CROSS-SECTION σ FOR $h\nu/mc^2 = E$

E	σ/r_0^2	$M''_0 + M''_1$	Δ'''	E	σ/r_0^2	$M''_0 + M''_1$
0.2	6.1739	1068	- 137	2.0	2.6315	78
0.25	5.8390	843	- 96	2.1	2.5688	69
0.3	5.5516	676	- 75	2.2	2.5097	64
0.35	5.3021	547		2.3	2.4539	59
0.4	5.0830	448		2.4	2.4012	53
0.45	4.8887	370		2.5	2.3513	48
0.5	4.7148	312		2.6	2.3039	44
0.55	4.5579	267		2.7	2.2588	41
0.6	4.4155	228		2.8	2.2158	38
0.65	4.2855	197		2.9	2.1748	35
0.7	4.1661	171		3.0	2.1356	126
0.75	4.0559	150		3.2	2.0622	110
0.8	3.9537	134		3.4	1.9946	97
0.85	3.8586	119		3.6	1.9322	85
0.9	3.7698	107		3.8	1.8743	75
0.95	3.6866	95		4.0	1.8204	68
1.0	3.6085	328		4.2	1.7700	61
1.1	3.4652	272		4.4	1.7229	55
1.2	3.3367	228		4.6	1.6786	50
1.3	3.2206	193		4.8	1.6370	46
1.4	3.1149	166		5.0	1.5977	243
1.5	3.0181	143		5.5	1.5084	198
1.6	2.9290	125		6.0	1.4300	163
1.7	2.8465	110		6.5	1.3605	135
1.8	2.7699	97		7.0	1.2984	113
1.9	2.6984	88		7.5	1.2424	98
2.0	2.6315			8.0	1.1916	86
				8.5	1.1454	74
				9.0	1.1032	62
				9.5	1.0644	50
				10	1.0284	

sum of adjacent second differences, modified by a ‘throw-back’ to simulate the effect of fourth differences. The error in $f(a + nh)$ due to using (14) is at most half a unit in the last tabulated place in σ/r_0^2 and one such unit in σ_a/r_0^2 .

TABLE A2. ENERGY ABSORPTION CROSS-SECTION σ_a FOR $h\nu/mc^2 = E$

E	σ_a/r_0^2	$M_0'' + M_1''$	Δ'''
0.4	1.1137	-535	144
0.5	1.1695	-331	
0.6	1.2038	-216	
0.7	1.2246	-148	
0.8	1.2363	-102	
0.9	1.2418	-73	
1.0	1.2430	-83	
1.125	1.2401	-57	
1.25	1.2339	-41	
1.375	1.2253	-28	
1.5	1.2150	-64	
1.75	1.1915	-25	
2	1.1659	11	
2.5	1.1134	62	
3	1.0625	337	
4	0.9711	271	
5	0.8940	217	
6	0.8289	174	
7	0.7735	142	
8	0.7258	110	
9	0.6846	74	
10	0.6479		

APPENDIX B. CALCULATION OF THE DOSE FROM RADIATION SCATTERED UP TO THREE TIMES

B1. *Properties of the Dirac delta-function*

It is convenient to list here some properties of the Dirac delta-function which we shall use without comment later in this appendix. The Dirac function $\delta(x)$ is defined by

$$\left. \begin{aligned} \delta(x) &= 0 \quad \text{if } x \neq 0 \\ \int_{-\infty}^{\infty} \delta(x) dx &= 1, \end{aligned} \right\} \tag{15}$$

and

so that $\delta(x)$ may be regarded as the derivative of a Heaviside step-function. Usually $\delta(x)$ appears multiplied by some well-behaved function, $f(x)$, say; we have then

$$\left. \begin{aligned} f(x) \delta(x) &= 0 \quad \text{if } x \neq 0 \\ \int_{-\infty}^{\infty} f(x) \delta(x) dx &= f(0), \end{aligned} \right\} \tag{16}$$

and

and it is only in such integrals that $f(x) \delta(x)$ enters any physical result. Following Dirac, we shall therefore not arrive at any inconsistent results if we proceed to write an equality sign between expressions involving delta-functions whenever we can show that they are both zero except at a certain value of x , and their integrals through this singular point are equal.

We shall need two equations involving delta-functions. The first is

$$\int_{-\infty}^{\infty} \delta(x-a) \delta(x-b) dx = \delta(a-b). \quad (17)$$

This is proved by noting that both sides are zero if $a \neq b$, and that

$$\int_{-\infty}^{\infty} da \int_{-\infty}^{\infty} \delta(x-a) \delta(x-b) dx = 1 = \int_{-\infty}^{\infty} \delta(a-b) da.$$

The second useful equation is

$$\delta(f(x) - f(x_0)) = \frac{\delta(x - x_0)}{|(df/dx)_{x_0}|}, \quad (18)$$

which follows from both sides being zero if $x \neq x_0$, and from

$$\begin{aligned} \int_{-\infty}^{\infty} \delta(f(x) - f(x_0)) dx &= \int_{-\infty}^{\infty} \frac{\delta(y - f(x_0)) dy}{|df/dx|} \quad \text{with } y = f(x) \\ &= 1 / \left| \left(\frac{df}{dx} \right)_{x_0} \right| = \int_{-\infty}^{\infty} \frac{\delta(x - x_0) dx}{|(df/dx)_{x_0}|}. \end{aligned}$$

B 2. The dose from once-scattered quanta

There is no difficulty in writing down the transformation process from n -times scattered radiation to the $(n+1)$ -times scattered intensity. It makes a clearer introduction to the subject, however, to derive the once-scattered radiation, then the twice-scattered component. The method for higher orders will be obvious. For simplicity we shall also assume in the derivations that the total cross-section from Compton scattering, pair-creation and photoelectric absorption is a constant, σ . We shall simply quote the rather more lengthy formulae which allow for the variation of the cross-section with $h\nu/mc^2$.

Let the incident quanta have $h\nu/mc^2 = \gamma$, number one per unit area of the slab, and arrive at perpendicular incidence. We define the radiation intensity $I(x, \mu, E)$ by taking a small area dA at a point in the slab a distance x from the front face, with its normal making an angle θ with the inward normal to the slab face; $I(x, \mu, E) dA dE d\Omega$ is the number of quanta which cross dA in directions lying in a small solid angle $d\Omega$ around the normal to dA , with quantum energies $h\nu/mc^2$ lying between E and $E + dE$. $\mu (= \cos \theta)$ has been used as a more convenient variable than θ .

The unscattered radiation can be represented by

$$I_0(x, \mu, E) = (1/\pi) e^{-N\sigma x} \delta(\mu - 1) \delta(E - \gamma). \quad (19)$$

Similarly, radiation which has been scattered once before arriving at depth x is described by an intensity $I_1(x, \mu, E)$ which contains only one delta-function, $\delta\{E - \gamma/(1 + \gamma(1 - \mu))\}$, since all values of μ from -1 to $+1$ are possible but at any μ the value of E is fixed by (3) of §2.2. Thus we can write

$$I_1(x, \mu, E) = I(x, \mu) \delta(E - \gamma/\{1 + \gamma(1 - \mu)\}). \quad (20)$$

Radiation scattered twice has an intensity I_2 which contains no delta-functions, and the same holds for I_3 and all higher orders.

Let $NF(\mu, E_1, E_2)d\Omega dE_2 ds$ be the probability that a quantum with $h\nu/mc^2 = E_1$ will be scattered during a travel ds into a solid angle $d\Omega$ with deflexion $\cos^{-1}(\mu)$ and final value of $h\nu/mc^2$ between E_2 and $E_2 + dE_2$. Since all scattered quanta have

$$E_2 = E_1/\{1 + E_1(1 - \mu)\},$$

we have, in fact,

$$F(\mu, E_1, E_2) = g(\mu, E_1) \delta(E_2 - E_1/\{1 + E_1(1 - \mu)\}), \quad (21)$$

where $g(\mu, E_1)$ is the usual scattering coefficient. It is often useful to transform (21) by the rule (18), giving

$$F(\mu, E_1, E_2) = g(\mu, E_1) \{1 + E_1(1 - \mu)\}^2 \delta(E_1 - E_2/\{1 - E_2(1 - \mu)\}). \quad (22)$$

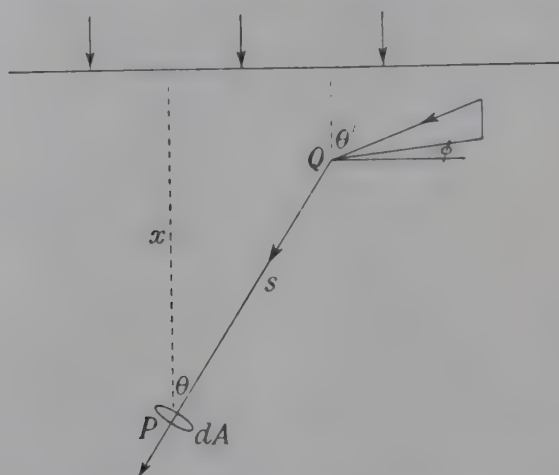


FIGURE 5. Notation for scattering of quanta which may have been scattered previously.

Consider now figure 5. Radiation arriving at Q from the direction (θ', ϕ') is scattered into the direction QP defined by θ . The deflexion in this scattering is θ_1 , where

$$\mu_1 = \cos \theta_1 = \mu\mu' + \sin \theta \sin \theta' \cos \phi',$$

with, as usual, $\mu = \cos \theta$, $\mu' = \cos \theta'$. As before, let the scattering volume subtend the solid angle $d\Omega$ at P and lie at distances between s and $s + ds$ from P . This volume, $d\tau = s^2 ds d\Omega$, is subjected to a field of unscattered radiation $I_0(x, \mu, E)$; that is, if the volume is given the form of a cylinder of cross-section dA' and length $d\tau/dA'$, the number of quanta with $h\nu/mc^2$ between E' and $E' + dE'$ which cross the area dA' in solid angle $d\Omega'$ is $I_0(x - s\mu, \mu', E') dE' dA' d\Omega'$, and of these the number which are scattered so as to hit dA at P and have $h\nu/mc^2$ between E and $E + dE$ is

$$\begin{aligned} I_0(x - s\mu, \mu', E') dE' dA' d\Omega' NF(\mu_1, E', E) dE \frac{dA}{s^2} \frac{d\tau}{dA'} e^{-N\sigma s} \\ = NdA dE d\Omega e^{-N\sigma s} I_0(x - s\mu, \mu', E') F(\mu_1, E', E) dE' d\Omega' ds. \end{aligned}$$

Taking for definiteness the case $\mu \geq 0$, we have by integration

$$I_1(x, \mu, E) = N \int_{E'=0}^{\infty} \int_{\Omega'} \int_{s=0}^{x/\mu} e^{-N\sigma s} I_0(x - s\mu, \mu', E') F(\mu_1, E', E) dE' d\Omega' ds. \quad (23)$$

Using (19) and (22), we have from (23)

$$I_1(x, \mu, E) = \frac{N}{\pi} \int_0^{x/\mu} e^{-N\sigma x - N\sigma s(1-\mu)} ds \int_{\mu'=-1}^1 \delta(\mu' - 1) d\mu' \int_{\phi'=0}^{2\pi} d\phi' \\ \times \int_{E'=0}^E g(\mu_1, E') [1 + E'(1 - \mu_1)]^2 \delta(E' - \gamma) \delta[E' - E/\{1 - E(1 - \mu_1)\}] dE',$$

which by (17) of § B 1 becomes

$$I_1(x, \mu, E) = \frac{N}{\pi} \int_0^{x/\mu} e^{-N\sigma x - N\sigma s(1-\mu)} ds \int_{\mu'=-1}^1 \int_{\phi'=0}^{2\pi} \delta(\mu' - 1) g(\mu_1, \gamma) \{1 + \gamma(1 - \mu_1)\}^2 \\ \times \delta[E/\{1 - E(1 - \mu_1)\} - \gamma] d\mu' d\phi'.$$

When $\mu' = 1$, $\mu_1 = \mu$; carrying out the integration over μ' ,

$$I_1(x, \mu, E) = \frac{N}{2\pi} \int_0^{x/\mu} e^{-N\sigma x - N\sigma s(1-\mu)} ds \\ \times \int_{\phi'=0}^{\infty} g(\mu, \gamma) \{1 + \gamma(1 - \mu)\}^2 \delta[E/\{1 - E(1 - \mu)\} - \gamma] d\phi'.$$

Integrating over ϕ' , and using (18), we end with

$$I_1(x, \mu, E) = g(\mu, \gamma) \delta[E - \gamma/\{1 + \gamma(1 - \mu)\}] e^{-N\sigma x} [1 - \exp\{-N\sigma x(1 - \mu)/\mu\}]/\sigma(1 - \mu) \quad (24)$$

for $\mu \geq 0$. When $\mu < 0$ we need to introduce the total thickness of the slab, D mean free paths, leading to

$$I(x, \mu) = g(\mu, \gamma) e^{-N\sigma x} [1 - \exp\{-(N\sigma x - D)(1 - \mu)/\mu\}]/\sigma(1 - \mu). \quad (25)$$

When the scattering cross-section σ depends on $h\nu/mc^2$ we write σ_1 for the value of σ at $h\nu/mc^2 = \gamma/\{1 + \gamma(1 - \mu)\}$, and σ_0 for $\sigma(\gamma)$, and then (24) and (25) must be replaced by

$$I(x, \mu) = g(\mu, \gamma) e^{-N\sigma_0 x} [1 - \exp\{-N(\sigma_1 - \mu\sigma_0)x/\mu\}]/(\sigma_1 - \mu\sigma_0) \quad (26)$$

for $\mu \geq 0$, or

$$I(x, \mu) = g(\mu, \gamma) e^{-N\sigma_0 x} [1 - \exp\{-(N\sigma_0 x - D)(\sigma_1 - \mu\sigma_0)/\mu\sigma_0\}]/(\sigma_1 - \mu\sigma_0) \quad (27)$$

for $\mu < 0$. The dose relative to that at the front face of the slab is

$$\text{Relative dose} = \int_{\mu=-1}^1 (E/\gamma) (\sigma_a(E)/\sigma_a(\gamma)) I(x, \mu) 2\pi d\mu \quad (28)$$

with

$$E = \gamma/\{1 + \gamma(1 - \mu)\}.$$

Introducing the differential cross-section for scattering of energy

$$f(\mu, \gamma) = g(\mu, \gamma)/\{1 + \gamma(1 - \mu)\},$$

and using the exact formula (26) we find for the relative dose at the far side of the slab, which is the most important place,

$$\text{Relative dose} = \frac{2\pi e^{-D}}{\sigma_a(\gamma)} \int_0^1 \frac{f(\mu, \gamma) \left[1 - \exp\left\{-\frac{D(\sigma_1 - \mu\sigma_0)}{\mu\sigma_0}\right\} \right] \sigma_a(E) d\mu}{(\sigma_1 - \mu\sigma_0)}. \quad (29)$$

In the modified first-order method we use (24) in place of (26) and replace $\sigma_a(E)$ by $1.243r_0^2$ if $\gamma \geq 1$ and otherwise by $\sigma_a(\gamma)$.

B3. *Twice-scattered radiation*

Equation (23) shows how to pass from unscattered to once-scattered radiation by a process which can obviously be applied to higher orders of scattering and which can be easily generalized to apply to negative μ and also to σ depending on $h\nu/mc^2$. We have already evaluated (23) when the intensity inside the integrand is a product of two delta-functions. To derive I_2 we have to integrate I_1 , which contains only one delta-function. For I_3 and higher orders there are no delta-functions and the process becomes the same for all orders. We derive here the formulae for twice-scattered radiation, assuming that $\mu \geq 0$ and that σ is independent of $h\nu/mc^2$. We have

$$I_2(x, \mu, E) = N \int_{E'=0}^{\infty} \int_{\Omega'} \int_{s=0}^{x/\mu} e^{-N\sigma s} I_1(x-s\mu, \mu', E') F(\mu_1, E', E) dE' d\Omega' ds \quad (30)$$

and

$$I_1(x-s\mu, \mu', E') = \frac{g(\mu', \gamma) e^{-N\sigma(x-s\mu)}}{\sigma(1-\mu')} \times [1 - \exp\{-N\sigma(x-s\mu)(1-\mu')/\mu'\}] \delta[E' - \gamma/(1+\gamma(1-\mu'))], \quad (31)$$

provided $\mu' \geq 0$. If $\mu' < 0$, the only effect is to replace $N\sigma(x-s\mu)$ in the exponential by $N\sigma(x-s\mu) - D$. In (30) we have delta-functions in E' arising from (22) and (31), and the integration over E' gives

$$I_2(x, \mu, E) = N \int_{\mu'=0}^1 \int_{\phi'=0}^{2\pi} \int_{s=0}^{x/\mu} \frac{e^{-N\sigma x - N\sigma s(1-\mu)} g(\mu', \gamma) g(\mu_1, \gamma/(1+\gamma(1-\mu')))}{\sigma(1-\mu') \{1-E(1-\mu_1)\}^2} \times [1 - \exp\{-N\sigma(1-\mu')(x-s\mu)/\mu'\}] \delta\left\{\frac{E}{1-E(1-\mu_1)} - \frac{\gamma}{1+\gamma(1-\mu')}\right\} d\mu' d\phi' ds, \quad (32)$$

together with an integral from $\mu' = -1$ to 0, in which the integrand differs only by an obvious change. Using (18) twice, (32) becomes

$$I_2(x, \mu, E) = N \int_{\mu'=0}^1 \int_{\phi'=0}^{2\pi} \int_{s=0}^{x/\mu} \frac{e^{-N\sigma x - N\sigma s(1-\mu)} g(\mu', \gamma) g(\mu_1, \gamma/(1+\gamma(1-\mu')))}{\sigma\gamma^2 \left|1 + \frac{d\mu_1}{d\mu'}\right| (1-\mu')} \times [1 - \exp\{-N\sigma(x-s\mu)(1-\mu')/\mu'\}] \{1+\gamma(2-\mu'-\mu_1)\}^2 \times \delta\left\{\mu' - \left(2-\mu_1 - \frac{1}{E} + \frac{1}{\gamma}\right)\right\} d\mu' d\phi' ds, \quad (33)$$

together with a very similar integral over negative μ' . The values of μ' and μ_1 which satisfy

$$2 - \mu_1 - \mu' = \frac{1}{E} - \frac{1}{\gamma} \quad (34)$$

depend only on E , γ , μ and ϕ' and may be written for short as μ'_0 and μ_{10} ; of course, they may not exist for certain ranges of these parameters. Integrating over μ' , (33) leads to

$$I_2(x, \mu, E) = N \int_{\phi'=0}^{2\pi} \int_{s=0}^{x/\mu} \frac{e^{-N\sigma x - N\sigma s(1-\mu)} g(\mu'_0, \gamma) g(\mu_{10}, \gamma/(1+\gamma(1-\mu'_0)))}{\sigma E^2 \left|1 + \mu - \mu'_0(\mu'_0 + \mu_{10})\right|} \times (1 + \mu'_0) [1 - \exp\{-N\sigma(x-s\mu)(1-\mu'_0)/\mu'_0\}] d\phi' ds, \quad (35)$$

provided $\mu'_0 \geq 0$; otherwise there is an obvious change in the integrand. Define M by

$$M = 1 \quad \text{if} \quad \mu'_0 \geq 0; \quad M = \exp\{D(1 - \mu'_0)/\mu'_0\} \quad \text{if} \quad \mu'_0 < 0.$$

The integration over s can be carried out and gives

$$I_2(x, \mu, E) = \sigma^{-2} \int_0^{2\pi} g(\mu'_0, \gamma) g(\mu_{10}, \gamma/1 + \gamma(1 - \mu'_0)) \frac{(1 + \mu'_0)}{E^2 |1 + \mu - \mu'_0(\mu'_0 + \mu_{10})|} \\ \times \left[\frac{e^{-N\sigma x}}{(1 - \mu)} [1 - \exp\{-N\sigma x(1 - \mu)/\mu\}] - \frac{M\mu'_0 e^{-N\sigma x/\mu'_0}}{(\mu'_0 - \mu)} \right. \\ \left. \times [1 - \exp\{-N\sigma x(\mu'_0 - \mu)/\mu\mu'_0\}] \right] d\phi', \quad (36)$$

provided $\mu \geq 0$ and σ is independent of $h\nu/mc^2$. When these restrictions are removed, we have, for $\mu \geq 0$,

$$I_2(x, \mu, E) = \int_0^{2\pi} g(\mu'_0, \gamma) g\left(\mu_{10}, \frac{\gamma}{1 + \gamma(1 - \mu'_0)}\right) \frac{(1 - \mu'^2_0)}{E^2 |1 + \mu - \mu'_0(\mu'_0 + \mu_{10})|} \frac{e^{-N\sigma_0 x}}{(\sigma_1 - \mu'_0 \sigma_0)} \\ \times \left[\frac{(1 - \exp\{(\sigma_2 - \mu\sigma_0)Nx/\mu\})}{(\sigma_2 - \mu\sigma_0)} - \frac{\mu'_0 M' \exp\{-(\sigma_1 - \mu'_0 \sigma_0)Nx/\mu'_0\}}{(\mu'_0 \sigma_2 - \mu\sigma_1)} \right. \\ \left. \times (1 - \exp\{-(\mu'_0 \sigma_2 - \mu\sigma_1)Nx/\mu'_0 \mu\}) \right] d\phi', \quad (37)$$

where M' is defined as $M' = 1$ if $\mu'_0 \geq 0$ and $M' = \exp\{(\sigma_1 - \mu'_0 \sigma_0)D/\mu'_0 \sigma_0\}$ if $\mu'_0 < 0$, and $\sigma_0, \sigma_1, \sigma_2$ are the values of σ at $h\nu/mc^2 = \gamma, E/\{1 - E(1 - \mu_{10})\}$, and E . For $\mu < 0$

$$I_2(x, \mu, E) = \int_0^{2\pi} g(\mu'_0, \gamma) g\left(\mu_{10}, \frac{\gamma}{1 + \gamma(1 - \mu'_0)}\right) \frac{(1 - \mu'^2_0)}{E^2 |1 + \mu - \mu'_0(\mu'_0 + \mu_{10})|} \frac{e^{-N\sigma_0 x}}{(\sigma_1 - \mu'_0 \sigma_0)} \\ \times \left[\frac{1}{(\sigma_2 - \mu\sigma_0)} (1 - \exp\{-(\sigma_2 - \mu\sigma_0)(N\sigma_0 x - D)/\mu\}) \right. \\ \left. - \frac{\mu'_0 M' \exp\{-(\sigma_1 - \mu'_0 \sigma_0)Nx/\mu'_0\}}{(\mu'_0 \sigma_2 - \mu\sigma_1)} \left\{ 1 - \exp\left\{-\frac{(\mu'_0 \sigma_2 - \mu\sigma_1)}{\mu'_0 \mu \sigma_0} (N\sigma_0 x - D)\right\} \right\} \right] d\phi'. \quad (38)$$

Various limiting forms are useful in computation but are hardly worth publication.

B4. Computation of twice-scattered radiation

For practical purposes we require the relative doses at the far side of slabs of various finite thicknesses, but to compute the dose due to thrice-scattered radiation we also require to know the intensity due to twice-scattered radiation, $I_2(x, \mu, E)$ at various depths.

The relative dose due to twice-scattered radiation at the underside of a block of thickness x is [cf. (28)]

$$\int_{E=0}^{\gamma} \int_{\mu=0}^1 (E/\gamma) (\sigma_a(E)/\sigma_a(\gamma)) I_2(x, \mu, E) 2\pi d\mu dE;$$

it is convenient therefore to compute $I_2(x, \mu, E)$ as a first step, considering only values of $\mu \geq 0$, at particular values of x for various values of μ and E . In the computation of relative dose it was found better to integrate first with respect to E rather than μ .

When $\mu = 1$ the expression for $I_2(x, \mu, E)$ simplifies greatly and can be integrated directly with respect to ϕ' . In consequence $I_2(x, 1, E)$ can be computed for a large number of values of E without undue labour.

When $\mu \neq 1$ it is necessary to determine the limiting values of E which are possible and also the valid range of ϕ' for any selected value of E within those limits.

The limits of E are found by solving the simultaneous equations

$$2 - \mu'_0 - \mu_{10} = \frac{1}{E} - \frac{1}{\gamma} = \xi, \text{ say} \quad (34)$$

and

$$\mu_{10} = \mu\mu'_0 + \sin \theta \sin \theta'_0 \cos \phi', \quad (39)$$

whose solution is

$$\mu'_0 = \frac{(1 + \mu)(2 - \xi) \pm \sqrt{(1 - \mu^2) \sqrt{[\cos^2 \phi' \{ \cos^2 \phi' (1 - \mu^2) + (1 + \mu)^2 - (2 - \xi)^2 \}]}}}{(1 + \mu)^2 + \cos^2 \phi' (1 - \mu^2)}. \quad (40)$$

The limiting values of E which give rise to real values of μ'_0 when $|\cos \phi'| = 1$, and the valid ranges of ϕ' for values of E within those limits, can now be found.

B5. Thrice-scattered radiation

With appropriate changes in notation an integral expression similar to (30) can be written down connecting $I_3(X, \mu_3, E_3)$, the intensity of thrice-scattered radiation at the underside of a block of thickness X , with $I_2(x, \mu_2, E_2)$ the intensity of twice-scattered radiation at depth x in the block.

Similarly, the relative dose due to thrice-scattered radiation at the far side of the block can be written [cf. (28)]

$$\iint (E/\gamma) (\sigma_a(E)/\sigma_a(\gamma)) I_3(X, \mu_3, E_3) 2\pi d\mu_3 dE_3.$$

From the computation of twice-scattered radiation we had $I_2(x, \mu_2, E_2)$ tabulated against x, μ_2 and E_2 ; in practice therefore it was much simpler to interpolate the required values of this function than to begin by computing $I_3(X, \mu_3, E_3)$ in terms of μ_3, μ_2, ϕ_2 etc. These values of $I_2(x, \mu_2, E_2)$ had in fact been computed for slabs of total thickness x , that is, without 'back-scatter'. This introduces an error, but from the results for the single-scatter dose computed with and without 'back-scatter' it is clear that this error is small, of the order of 5 % at most, for γ greater than 2.

If the ultimate object were to compute the dose due to fourth and higher scattered radiation it would be necessary to compute $I_3(X, \mu_3, E_3)$ as the first step in order to be able to repeat the process described above. In our work, however, only the contribution up to thrice-scattered radiation was required. In consequence we had a wider choice in the order of integration of the expression for the relative dose; this permitted the valid ranges of the variables of integration to be determined with relatively little labour and also enabled the interpolated values of $I_2(x, \mu_2, E_2)$ to be calculated with the maximum accuracy.

When $\mu_3 = 1$ or $\mu_2 = 1$ the expression for relative dose can be integrated at once with respect to ϕ_2 .

APPENDIX C. AN UPPER LIMIT TO THE DOSE

C1. Formulation of an upper limit

Let the relative dose at the far side of a slab D mean free paths thick be $e^{-\alpha D}$, where α is a function of D and of $h\nu/mc^2 = \gamma$. Of this dose, e^{-D} arises from radiation which has passed through without being scattered and $e^{-\alpha D} - e^{-D}$ comes from radiation scattered at least once. We seek an upper limit for the dose, that is, a lower limit to α .

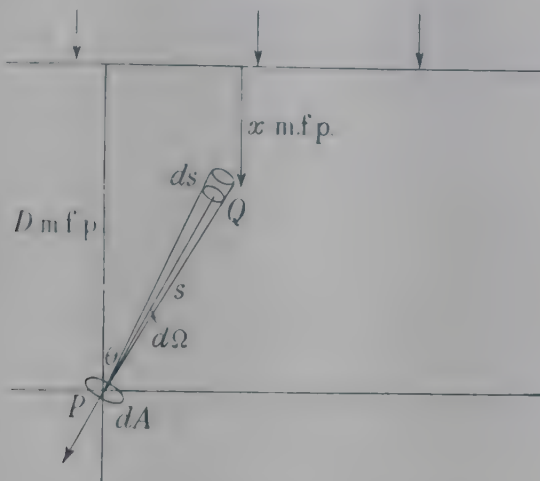


FIGURE 6. Notation for upper limit theory.

Consider (figure 6) scattered radiation crossing an area dA at a point P at the far side of the slab, with directions in a small solid angle $d\Omega$ around the direction defined by θ . We classify this radiation by the point at which its last scattering occurred. The volume element round a point Q which is x mean free paths from the surface, with distances from P between s and $s + ds$ and subtending $d\Omega$ at P has volume $d\tau = s^2 ds d\Omega$ and scatters across dA at P from the *unscattered* radiation at Q

$$N d\tau g(\mu, \gamma) e^{-x - N\sigma(E)s} \frac{dA}{s^2}$$

quanta with $h\nu = E mc^2$, where

$$E = \gamma / \{1 + \gamma(1 - \mu)\},$$

and so they contribute to the relative dose at P an amount

$$\{\sigma_a(E)/\sigma_a(\gamma)\} N f(\mu, \gamma) e^{-x - N\sigma(E)s} ds d\Omega.$$

The functions f and g were defined in (1) and (4) of § 2.2.

We also have scattering into direction QP from the scattered radiation passing Q . This radiation gives at Q a relative dose $e^{-x\alpha(x)} - e^{-x}$; the dependence of α on γ need not be shown explicitly. Let this radiation be assigned an average quantum energy $E' mc^2$ and an inclination whose cosine is μ' , where E' and μ' are to be chosen to maximize the effect at P . E' and μ' can be functions only of γ and x .

The number of quanta per unit area in the scattered radiation at Q is then

$$\gamma \sigma_a(\gamma) \{e^{-x\alpha(x)} - e^{-x}\} / E' \sigma_a(E'),$$

and these can be scattered from μ' to μ (with a relative azimuthal angle ϕ) by a deflexion whose cosine is

$$\mu_1 = \mu\mu' + \sin\theta \sin\theta' \cos\phi,$$

so that the number of these quanta scattered by the $Nd\tau$ electrons at Q to reach dA at P is

$$\int_{\phi=0}^{2\pi} Ndsd\Omega g(\mu_1, E') \gamma \sigma_a(\gamma) [e^{-x\alpha(x)} - e^{-x}] e^{-N\sigma(E'')s} d\phi / 2\pi E' \sigma_a(E'),$$

where

$$E'' = E' / \{1 + E'(1 - \mu_1)\}.$$

The contribution to the relative dose at P is

$$\int_{\phi=0}^{2\pi} E''' \sigma_a(E''') Ng(\mu_1, E') [e^{-x\alpha(x)} - e^{-x}] e^{-N\sigma(E'')s} ds d\Omega d\phi / 2\pi E' \sigma_a(E').$$

Hence

$$e^{-D\alpha(D)} - e^{-D} = \frac{N}{\sigma_a(\gamma)} \iint \sigma_a(E) f(\mu, \gamma) e^{-x-N\sigma(E)s} ds d\Omega \\ + N \iiint \sigma_a(E'') f(\mu_1, E') \{e^{-x\alpha(x)} - e^{-x}\} e^{-N\sigma(E'')s} ds d\Omega d\phi / 2\pi \sigma_a(E'). \quad (41)$$

It will have been noticed that $e^{-D\alpha(D)}$ is used both for the dose at the far side of a slab of thickness D mean free paths and for the dose at depth D inside a block of unspecified and possibly very large thickness. The error will be small so long as back-scattering effects are negligible.

Write σ_1 for $\sigma(E)$ and σ_0 for $\sigma(\gamma)$. The first integral on the right of (41) is

$$\frac{2\pi e^{-D}}{\sigma_a(\gamma)} \int_0^1 \sigma_a(E) f(\mu, \gamma) \{e^{D(1-\sigma_1/\mu\sigma_0)} - 1\} d\mu / (\mu\sigma_0 - \sigma_1),$$

and the second integral on the right of (41) can be written

$$\int_0^D \{e^{-x\alpha(x)} - e^{-x}\} dx \int_0^1 \frac{d\mu}{\mu} \int_0^{2\pi} \frac{\sigma_a(E'') f(\mu_1, E') e^{-(D-x)\sigma(E'')/\sigma_0\mu} d\phi}{\sigma_0 \sigma_a(E')}, \quad (42)$$

where E' and μ' are functions only of x and γ , and are chosen to maximize the double integral over μ and ϕ in (42); that is, for a given dose from scattered radiation at depth x we must choose its direction and wave-length to maximize the dose received at depth D after one more scattering at depth x . An indication of the answer is provided by our computations of singly-scattered radiation received behind a slab, for perpendicular incidence. The two problems differ, of course, in that the latter one includes an averaging over all depths for the scattering point. It is found that for small D , say 0.3, it pays to use a small value of $h\nu/mc^2$ for then the increased probability of large-angle scatter more than counterbalances the larger total scattering cross-section. At D of order unity the optimum $h\nu/mc^2$ is the maximum, γ . We conclude that E' should be replaced by γ unless $D - x$ is small, of order 0.3. If we replace E' by γ at all depths x we cannot expect to get an upper limit to the dose unless D is more than, say, two or three.

This argument is based on radiation at vertical incidence. It can, however, be shown from the same computations that if $D - x$ is less than about 1 there is a little to be gained by inclining the radiation. This would complicate the theory, so we use $\mu' = 1$ at all values of x , thereby renouncing a true upper limit until D approaches 3,

Let us, then, regard the dose at x as due entirely to radiation with $h\nu/mc^2 = \gamma$, travelling normal to the slab face. In place of (41) we have

$$e^{-D\alpha(D)} - e^{-D} = \frac{2\pi}{\sigma_a(\gamma)} \int_0^1 \sigma_a(E) f(\mu, \gamma) d\mu \int_0^{D/N\sigma_0\mu} e^{-N\sigma_1 s - x\alpha(x)} N ds. \quad (43)$$

C2. Cut-off effects

Assume that $\alpha(x)$ can be taken to be independent of x for small x . Equation (43) gives, for small D ,

$$e^{-D\alpha} - e^{-D} = \frac{2\pi}{\sigma_a(\gamma)} \int_0^1 \frac{\sigma_a(E) f(\mu, \gamma) (e^{-D\alpha} - e^{-D\sigma_1/\mu\sigma_0}) d\mu}{(\sigma_1 - \sigma_0\mu\alpha)}, \quad (44)$$

and we might expect to be able to solve for α by letting D tend to zero. This is prevented by the fact that for any assigned D we can, by going to sufficiently small μ , prevent the expansion of $e^{-D\sigma_1/\mu\sigma_0}$ in a power series. Suppose, therefore, we cut off μ at a small value, μ_0 . Then α from (44) is almost independent of μ_0 for D greater than 0.1. As D is reduced, tending to zero, α decreases, the more rapidly the smaller μ_0 . For sufficiently small μ_0 , α can even go negative. This is of course quite possible for an upper limit, though it is not very helpful. Limit values of α as D tends to zero are, for example, with $\gamma = 10$, $\alpha = 0.45$ for $\mu_0 = 0.3$ and $\alpha = 0.30$ for $\mu_0 = 0.02$.

In any practical case there will be a cut-off at finite μ , since one never does have an unsupported slab extending to infinity in its own plane. If it is assumed, for example, that supports or walls block out gamma rays from parts of the slab more than 10 ft. away, and that we are not interested in distances less than 6 in. from the slab, μ_0 would be 0.05. In any case, these difficulties occur only for slabs less than 0.1 mean free path thick.

C3. Computation of the upper limit

For sufficiently small D , the value of $\alpha(D)$ from (43) is negative. This certainly gives an upper limit to the dose, but $\alpha = 0$ is even better. Suppose $\alpha = 0$ for $D \leq D_0$; then (43) becomes

$$1 - e^{-D_0} = 2\pi \int_0^1 \sigma_a(E) f(\mu, \gamma) [1 - e^{-D_0\sigma_1/\mu\sigma_0}] d\mu / \sigma_1 \sigma_0(\gamma). \quad (45)$$

This can be solved for D_0 by trial. For $\gamma = 10$, D_0 is approximately 10^{-5} .

For larger D we replace the integral equation (43) by an integro-differential equation which can be solved in a step-by-step integration. (43) is

$$e^{-D\alpha(D)} - e^{-D} = \frac{2\pi}{\sigma_0 \sigma_a(\gamma)} \int_0^1 \sigma_a(E) f(\mu, \gamma) \frac{d\mu}{\mu} \int_0^D e^{-x\alpha(x) - \sigma_1(D-x)/\mu\sigma_0} dx, \quad (46)$$

which gives

$$\begin{aligned} \frac{d}{dD} (D\alpha(D)) - e^{-D(1-\alpha(D))} &= -\frac{2\pi}{\sigma_0 \sigma_a(\gamma)} \int_0^1 \sigma_a(E) f(\mu, \gamma) \\ &\times \left[1 - \frac{e^{D\alpha(D)} \sigma_1}{\sigma_0 \mu} \int_0^D e^{-x\alpha(x) - \sigma_1(D-x)/\mu\sigma_0} dx \right] \frac{d\mu}{\mu}. \end{aligned} \quad (47)$$

When $\alpha(D) \ll 1$, we have

$$\frac{d}{dD}(D\alpha(D)) \simeq e^{-D} - \frac{2\pi}{\sigma_0 \sigma_a(\gamma)} \int_0^1 \sigma_a(E) f(\mu, \gamma) e^{-D\sigma_1/\mu\sigma_0} d\mu/\mu, \quad (48)$$

which is exact at $D = D_0$.

For $\gamma = 10$, (48) was used from $D = D_0 = 10^{-5}$ to $D = 0.1$. The resulting values of $\alpha(D)$ were used in an alternative form of (47),

$$\begin{aligned} \frac{d}{dD}(D\alpha(D)) - e^{-D+D\alpha(D)} = & -\frac{2\pi}{\sigma_0 \sigma_a(\gamma)} \int_0^1 \sigma_a(E) f(\mu, \gamma) e^{-D\sigma_1/\mu\sigma_0} \\ & \times \left[1 + \int_0^D \frac{\sigma_1}{\mu\sigma_0} e^{\sigma_1 x/\mu\sigma_0} \{1 - e^{D\alpha(D) - x\alpha(x)}\} dx \right] \frac{d\mu}{\mu}. \end{aligned} \quad (49)$$

The difference between (48) and (49) is 9 % at $D = 0.1$ (and $\gamma = 10$) and only 10 % of this difference arises from the terms inside the integral sign in (49). Thus it was easy to correct to the exact values of $\alpha(D)$ over the range $D = 10^{-5}$ to 0.1. Thereafter integration was carried out by step-by-step methods from $D = 0.1$ to 0.25 in steps of 0.05. From $D = 0.25$ it was found to be more convenient to compute from

$$\begin{aligned} \frac{d}{dD}(D\alpha(D)) = & e^{-D+\alpha(D)} - \frac{2\pi}{\sigma_0 \sigma_a(\gamma)} \int_0^1 \sigma_a(E) f(\mu, \gamma) \\ & \times \left[e^{-D\sigma_1/\mu\sigma_0} - \frac{\sigma_1}{\mu\sigma_0} \int_0^D \{e^{D\alpha(D) - x\alpha(x)} - 1\} e^{-\sigma_1(D-x)/\mu\sigma_0} dx \right] \frac{d\mu}{\mu}. \end{aligned} \quad (50)$$

The integration was then carried on in the normal way, gradually increasing the length of step in D . At $D = 6.5$ the value of α began to increase. This is not likely to be the behaviour of the exact α , though it does not prevent the solution still giving an upper limit to the dose. We can take the value of α at $D = 6.5$ as giving an upper limit to the dose (lower limit to α) for all larger thicknesses. This limit is $\alpha = 0.753$ for $\gamma = 10$.

No doubt by including double scattering explicitly an extension of (43) could be derived whose α would not flatten out until larger values of D . However, there would probably be a similar extension of the region of small D in which the solution is not a true upper limit. In any case, the labour of computation has not encouraged further work. The work involved in this version, in which only single-scattering is taken into account explicitly, is roughly the same as the computation of the exact twice-scattered dose.

In (49) and (50) computation becomes difficult near $\mu = 0$, and this region was filled by an analytical integration.

C4. The 'modified upper limit' solution

We can get an approximate solution, which turns out to be extremely good, by reverting to (43) and replacing $\alpha(x)$ by $\alpha(D)$ inside the integral. This gives (44), namely,

$$e^{-D\alpha(D)} - e^{-D} = \frac{2\pi}{\sigma_a(\gamma)} \int_0^1 \frac{\sigma_a(E) f(\mu, \gamma) (e^{-D\alpha(D)} - e^{-D\sigma_1/\mu\sigma_0})}{\sigma_1 - \mu\sigma_0 \alpha} d\mu.$$

For given γ and D this equation can be solved by tabulating both sides over a small range of α .

As was shown in § 7, this approximation gives a close fit to the 'upper limit' in the region of small D , where the latter is not a true upper limit but is still a good approximation to the truth. For D larger than 4 the 'modified upper limit' gives a dose which becomes increasingly smaller than the rigorous upper limit, and in fact the 'modified upper limit' appears to lie very close to the truth at all thicknesses although for D larger than, say, 15, there is very little evidence against which to test the 'modified upper limit', whose behaviour, however, continues to be plausible.

The agreement between the 'modified' and 'rigorous' upper limit solutions arises at small D because the value of α has little effect on the solution here; as we have said in the preceding section, even putting $\alpha = 0$ inside the integral of (43) gives only a few per cent error. It matters little, therefore, that $\alpha(D)$ really varies quite rapidly with D at small D . On the other hand, at large D , where the exact variation of α is important, this variation is already slow and α can be taken to be constant without much error.

APPENDIX D. THE 'EXPONENTIAL APPROXIMATION WITHOUT DEGRADATION'

$$\text{For } \mu \text{ near 1, we have } g(\mu, \gamma) \simeq r_0^2 e^{-2\gamma(1-\mu)} \quad (51)$$

$$\text{and } f(\mu, \gamma) \simeq r_0^2 e^{-3\gamma(1-\mu)}. \quad (52)$$

These are exact to terms of order $1 - \mu$ in an expansion in powers of $1 - \mu$, and are qualitatively correct in showing a rapid decrease of scattering as the deflexion is increased. Indeed this property is rather too marked in these approximations, as is shown by table D 1.

TABLE D 1. ACCURACY OF THE APPROXIMATION TO $f(\mu, \gamma)$

For $\gamma = 10$.

μ	$f(\mu, \gamma)/r_0^2$, exact	from (52)
1	1	1
0.98	0.577	0.549
0.94	0.257	0.165
0.9	0.144	0.067
0.7	0.029	0.0001

In the approximate theory to be developed here we shall neglect the change of energy during scattering. If this were really correct f and g would be identical, for f gives the probability of energy being scattered and g the probability for scattering of quanta. To allow for this uncertainty and to correct to some extent for the insufficient scattering at large angles, we replace $f(\mu, \gamma)$ and $g(\mu, \gamma)$ by the scattering probability

$$P(\mu, \gamma) = r_0^2 e^{-\beta(1-\mu)}, \quad (53)$$

where β is expected to be of order 2γ . We shall assume $\beta \gg 1$. Since we shall solve our equations only to order $1 - \mu$ we gain nothing by multiplying the right side of (53) by another arbitrary parameter.

Let $I(x, \mu)$ be the intensity at depth x from the front face of a semi-infinite block. The radiation incident on the slab is taken to be one quantum per unit area, with $h\nu/mc^2 = \gamma$, and arrives at normal incidence.

As the degradation of quanta to smaller values of $h\nu$ is neglected, we can write $P(\mu, \gamma)$ simply as $P(\mu)$. The transfer equation can be set up in the usual way and is

$$\mu \frac{\partial}{\partial x} I(x, \mu) = -N\sigma I(x, \mu) + N \int_{-1}^1 \int_0^{2\pi} I(x, \mu_1) P(\mu\mu_1 + \sin\theta \sin\theta_1 \cos\phi_1) d\mu_1 d\phi_1, \quad (54)$$

where σ is the total cross-section for absorption and scattering of gamma rays with $h\nu/mc^2 = \gamma$. We leave σ as an arbitrary parameter, and write

$$I(x, \mu) = (1/\pi) e^{-N\sigma x} \delta(\mu - 1) + J(x, \mu), \quad (55)$$

so that J represents the radiation which has been scattered at least once, and satisfies

$$\mu \frac{\partial J}{\partial x} + N\sigma J(x, \mu) = N e^{-N\sigma x} P(\mu) + N \int_{-1}^1 \int_0^{2\pi} J(x, \mu_1) P(\mu\mu_1 + \sin\theta \sin\theta_1 \cos\phi_1) d\mu_1 d\phi_1. \quad (56)$$

We assume a solution

$$J(x, \mu) = (\beta/2\pi) e^{-N\sigma x} A(x) e^{-\beta(1-\mu)B(x)}, \quad (57)$$

where $A(x)$ and $B(x)$ are functions to be determined from (56). We shall discuss this assumption in detail a little later. Meanwhile we shall examine its consequences. For small x , the only ingoing scattered radiation is single-scattered, so that

$$B(0) = 1, \quad (58)$$

and at the surface there is no ingoing scattered radiation, so that

$$A(0) = 0. \quad (59)$$

There is, of course, some outgoing scattered radiation which is not represented by (57)–(59). This we neglect.

$$\text{We define a parameter } \lambda \text{ by} \quad \lambda = 2\pi r_0^2/\sigma\beta, \quad (60)$$

$$\text{and write} \quad y = N\lambda\sigma x. \quad (61)$$

Let dashes denote differentiation with respect to y . Using (53) and (57), (56) becomes

$$\begin{aligned} \mu A' - \mu(1-\mu)\beta AB' + A(1-\mu)/\lambda &= e^{-\beta(1-B)(1-\mu)} \\ &+ \beta A \int_{-1}^1 e^{\beta B(\mu_1-\mu)+\beta(\mu\mu_1-1)} I_0(\beta \sin\theta \sin\theta_1) d\mu_1, \end{aligned} \quad (62)$$

where I_0 is a Bessel function

$$I_0(a) = 1 + a^2/4 + a^4/64 + \dots$$

When $\mu = 1$, the equation of transfer becomes

$$A' = 1 + A/(1+B), \quad (63)$$

where we have used our assumption that $\beta \gg 1$. We get another relation between A and B by satisfying the transfer equation to first-order terms in $1-\mu$, leading to

$$\beta A \{ B' + (B-1)/A + B^2/(1+B)^2 \} + A/(1+B) + A/(1+B)^2 - A/(1+B)^3 + 1 - A/\lambda = 0. \quad (64)$$

If $\beta A \gg 1$, we need keep only the terms inside the brackets, and the solution of (63) and (64) becomes

$$A = y + y^2/4 + 5y^3/96 + 55y^4/6144 + 323y^5/245760 + 997y^6/5898240 + \dots, \quad (65)$$

$$B = 1 - y/8 + 7y^3/12288 + 3y^4/81920 - 11y^5/3932160 - 167y^6/293601280 + \dots \quad (66)$$

For $y \gg 1$,

$$A \sim 2e^y/(1+y), \quad (67)$$

$$B \sim 1/y - \frac{1}{2}y e^{-y}. \quad (68)$$

The relative dose at depth x is

$$2\pi \int_{-1}^1 I(x, \mu) d\mu = e^{-N\sigma x}(1 + A/B). \quad (69)$$

If $\beta A \gg 1$, the dose is

$$e^{-N\sigma x} [1 + y + 3y^2/8 + 19y^3/192 + 85y^4/4096 + 899y^5/245760 + 1099y^6/1966080 + \dots] \quad (70)$$

for small y , and for $y \gg 1$ the dose is

$$\sim 2 e^{-N\sigma x(1-\lambda)}. \quad (71)$$

Allowance for higher terms in (70) can be made by multiplying the y^6 term by

$$1 + 0.136y + y^2/60(1 - 0.12y),$$

which gives the dose for y up to 5 with an error not greater than 10 %.

If βy is comparable to unity our solution does not satisfy (64). This is due to our choice of the approximate form (57), with its boundary conditions (58) and (59) which describe correctly the ingoing but not the outgoing scattered radiation. We can, however, claim that our solution is valid whenever the scattered radiation is comparable with the unscattered component. (64) is not satisfied for βy of order unity, that is, where the scattered radiation is mainly outgoing, but the dose (70) has the correct behaviour in this region and cannot be more than a few per cent in error. These difficulties are confined to the first inch of a concrete block.

The effective semi-angle of the scattered radiation at depth x is given by

$$\beta B(x)(1 - \mu) \sim 3$$

for at this deflexion the intensity has dropped to about five per cent of its axial value. This angle opens out very slowly as we go into the slab.

To test this theory, σ and β were taken as arbitrary parameters and the dose: depth curve was compared with other solutions for $\gamma = 10$. If we write the relative dose as $e^{-D\alpha(D)}$, then for most values of σ and β it is found that α decreases over a wide range of increasing values of D . This is certainly the wrong behaviour and can be avoided only by accepting a gross error in α at all thicknesses.

APPENDIX E. THE 'EXPONENTIAL APPROXIMATION WITH DEGRADATION OF ENERGY'

E1. An equation of transfer

The disappointing results obtained from the theory of appendix D suggested an attempt to include in the theory the degradation of quanta which occurs in Compton scattering. The equations are naturally much more complicated. It is now harder

to find an approximate form for the intensity, which depends on three variables; even when an approximation has been found it is difficult to solve the transfer equation to determine the unknown functions. It cannot be claimed that a really satisfactory result has been reached. We have obtained a solution for small deflexions and energy degradations and have set up a function which is a reasonable extrapolation to larger deflexions and which gives a relative dose in terms of simple integrals. However, it appears that a major part of the dose comes from large deflexions and energy degradations, for which the function assumed is rather arbitrary. The numerical results have been discussed in §7. They can be made to give roughly the right trend of $\alpha(D)$ with D , but no close agreement with the true curve has been obtained. It is clear that it would be difficult to guarantee good results from an analogous method in problems with a more complicated geometry.

Let $\sigma(E)$ be the total cross-section for absorption and scattering of quanta with $h\nu/mc^2 = E$. We shall write σ for $\sigma(\gamma)$, where γ is the value of $h\nu/mc^2$ for the incident quanta. Let $I(x, \mu, E)$ be the intensity, defined as in §B 2 in terms of the number of quanta crossing a small area. Let ϕ be an azimuthal angle round the normal to the face of the block, which is taken to have a very large depth.

The transfer equation is

$$\mu \frac{\partial}{\partial x} I(x, \mu, E) + N\sigma(E) I(x, \mu, E) = N \iint I(x, \mu', E/\{1 - E(1 - \mu_1)\}) \times g(\mu_1, E/\{1 - E(1 - \mu_1)\}) d\mu' d\phi' / \{1 - E(1 - \mu_1)\}^2, \quad (72)$$

where

$$\mu_1 = \mu\mu' + \sin \theta \sin \theta' \cos \phi', \quad (73)$$

and the Compton scattering coefficient g is explained in §2.2. The unscattered radiation at depth x is

$$I_0(x, \mu, E) = (1/\pi) e^{-N\sigma x} \delta(\mu - 1) \delta(E - \gamma),$$

and the once-scattered component is

$$\begin{aligned} I_1(x, \mu, E) &= \frac{g(\mu, \gamma) e^{-N\sigma x}}{(\sigma_1 - \mu\sigma)} [1 - \exp\{-N(\sigma_1 - \mu\sigma)x/\mu\}] \delta(E - \gamma/\{1 + \gamma(1 - \mu)\}) \\ &\quad \text{for } \mu \geq 0, \\ &= \frac{g(\mu, \gamma) e^{-N\sigma x}}{(\sigma_1 - \mu\sigma)} \delta(E - \gamma/\{1 + \gamma(1 - \mu)\}) \quad \text{for } \mu < 0, \end{aligned}$$

where σ_1 is $\sigma(E)$ at

$$E = \gamma/\{1 + \gamma(1 - \mu)\}.$$

If we write

$$I = I_0 + I_1 + J, \quad (74)$$

we are able to work with a function J which is free from delta-functions in μ and E . The transfer equation for J is

$$\begin{aligned} &\left\{ \mu \frac{\partial}{\partial x} + N\sigma(E) \right\} J(x, \mu, E) \\ &= N \iint J(x, \mu', E/\{1 - E(1 - \mu_1)\}) g(\mu_1, E/\{1 - E(1 - \mu_1)\}) d\mu' d\phi' / \{1 - E(1 - \mu_1)\}^2 \\ &\quad + N e^{-N\sigma x} \int_0^{2\pi} \frac{g(\mu'_0, \gamma) g(\mu_{10}, \gamma/\{1 + \gamma(1 - \mu'_0)\})}{(\sigma_1 - \mu'_0\sigma) E^2 |1 + (d\mu_1/d\mu')_0|} [1 - \exp\{-N(\sigma_1 - \mu'_0\sigma)x/\mu'_0\}] d\phi', \end{aligned} \quad (75)$$

where $\sigma_1 = \sigma(\gamma/\{1 + \gamma(1 - \mu'_0)\})$, and μ'_0 and μ_{10} are the solutions of

$$\mu'_0 + \mu_{10} = 2 - \xi, \quad (76)$$

with
$$\mu_{10} = \mu\mu'_0 + \sin\theta \sin\theta'_0 \cos\phi', \quad (77)$$

where
$$\xi = \frac{1}{E} - \frac{1}{\gamma} \quad (78)$$

is a convenient measure of the energy degradation of a quantum with $h\nu/mc^2 = E$, relative to the incident quanta with $h\nu/mc^2 = \gamma$.

If $\mu'_0 \leq 0$ one omits the exponential term inside the second integral in (75).

The origin of the various terms in (75) will be clear from the discussion of twice-scattered radiation in § B 3. If the first integral on the right of (75) were omitted the solution J would just be the twice-scattered intensity I_2 . The first integral supplies the intensity from scatterings of higher order.

We introduce a function $L(x, \mu, \xi)$, defined by

$$L(x, \mu, \xi) = E^2 J(x, \mu, E), \quad (79)$$

which from (75) satisfies the transfer equation

$$\begin{aligned} \left\{ \mu \frac{\partial}{\partial x} + N\sigma(E) \right\} L(x, \mu, \xi) = N \iint L(x, \mu', \xi - 1 + \mu_1) g(\mu_1, E/\{1 - E(1 - \mu_1)\}) d\mu' d\phi' \\ + N e^{-N\sigma x} \int_0^{2\pi} \frac{(1 - \mu_0'^2) g(\mu_0', \gamma) g(\mu_{10}, \gamma/\{1 + \gamma(1 - \mu_0')\})}{(\sigma_1 - \mu_0'\sigma) |1 + \mu - \mu_0'(2 - \xi)|} \\ \times [1 - \exp\{-N(\sigma_1 - \mu_0'\sigma)x/\mu_0'\}] d\phi', \quad (80) \end{aligned}$$

with the exponential omitted inside the second integral if $\mu'_0 < 0$. Equation (80) is exact.

E2. Approximate solution

We replace the scattering coefficient g by the approximation

$$g(\mu, E) \simeq r_0^2 e^{-2\nu E(1-\mu)}, \quad (81)$$

which is exact to terms of order $1 - \mu$ for all E if $\nu = 1$. As the large-angle scattering is considerably underestimated by (81), one may expect that the arbitrary parameter ν will turn out to be less than unity. It can be shown, by considering the total amount of energy scattered, that a value of ν as small as 0.15 is not unreasonable when $\gamma = 10$.

The transfer equation becomes

$$\begin{aligned} \left\{ \mu \frac{\partial}{\partial x} + N\sigma(E) \right\} L(x, \mu, \xi) = Nr_0^2 \iint L(x, \mu', \xi - 1 + \mu_1) \exp \left[\frac{-2\nu(1 - \mu_1)}{\frac{1}{\gamma} + \xi - (1 - \mu_1)} \right] d\mu' d\phi' \\ + Nr_0^4 e^{-N\sigma x} \int_0^{2\pi} \frac{(1 - \mu_0'^2)}{(\sigma_1 - \mu_0'\sigma) |1 + \mu - \mu_0'(2 - \xi)|} \\ \times \exp \left[\frac{-2\gamma\nu\{\xi + \gamma(1 - \mu_0'^2)\}}{1 + \gamma(1 - \mu_0')} \right] [1 - \exp\{-N(\sigma_1 - \mu_0'\sigma)x/\mu_0'\}] d\phi', \quad (82) \end{aligned}$$

with the usual modification in the second integrand if $\mu'_0 < 0$.

We write $\eta = 1 - \mu, \quad s^2 = \xi^2 + \eta^2, \quad t = \xi/\eta, \quad (83)$

and attempt to solve (82) for small ξ and η . There is a singularity of L at $\xi = \eta = 0$, that is, for zero deflexion and energy degradation; we can, however, write

$$L(x, \mu, \xi) = M(x, t) + sP(x, t) + O(s^2) \quad (84)$$

for small s . It will appear later that (84) has an acceptable form near $\xi = \eta = 0$. We proceed to evaluate the various parts of (82) in terms of M and P . In the first place,

$$\left\{ \mu \frac{\partial}{\partial x} + N\sigma(E) \right\} L(x, \mu, \xi) = \frac{\partial M}{\partial x} + N\sigma M + s \left[\frac{\partial P}{\partial x} + N\sigma P + \frac{NMt}{\sqrt{(1+t^2)}} \left\{ \frac{\partial \sigma}{\partial \xi} \right\}_{\xi=0} - \frac{1}{\sqrt{(1+t^2)}} \frac{\partial M}{\partial x} \right] + O(s^2). \quad (85)$$

We assume $2\gamma\xi \ll 1$, and $N\sigma x$, the distance into the block in mean free paths for the incident radiation, less than or comparable with 2γ ; then $N\sigma x\xi \ll 1$ and $N(\sigma_1 - \mu'_0\sigma)x/\mu'_0 \ll 1$. The second integral in (82) becomes

$$N^2 r_0^4 x e^{-N\sigma x} (1 - 2\gamma\nu\xi) \int_0^{2\pi} (1 - \mu_0'^2) d\phi' / \mu'_0 |2 - \eta - \mu'_0(2 - \xi)|. \quad (86)$$

It is clear that the integral in (86) can be expressed as a function of t together with s times a function of t , and higher powers of s , and that the ratio of the s term to the first term will be of order unity. The s term in the expansion of the whole of (86) must arise mainly from the factor $1 - 2\gamma\nu\xi$ outside the integral. Hence we need to evaluate the integral in (86) only as far as zero-order terms in ξ and η .

We have $\mu'_0 = 1 + O(\eta)$, which reduces (86) to

$$4N^2 r_0^4 x (1 - 2\gamma\nu\xi) e^{-N\sigma x} \int_0^\pi (1 - \mu_0') d\phi' / |2 - \eta - \mu'_0(2 - \xi)|. \quad (87)$$

Also
$$(1 - \mu'_0) / |2 - \eta - \mu'_0(2 - \xi)| = \left| 2 - \eta + \frac{(t-1)\mu'_0\eta}{(1-\mu'_0)} \right|^{-1} \\ = \left(\frac{1}{2} \right) \left| 1 - \frac{t\eta}{2} + \frac{(t-1)\eta}{2(1-\mu'_0)} \right|^{-1} \\ = \frac{1}{2} \left\{ 1 + \frac{(t-1)\eta}{2(1-\mu'_0)} \right\}^{-1} + O(\eta). \quad (88)$$

The exact solution of (76) and (78) is

$$\mu'_0 = \frac{4 - 2\eta - 2\xi + \xi\eta + \cos\phi' [(2\eta - \eta^2) \{4\xi - 4\eta + \eta^2 - \xi^2 + (2\eta - \eta^2) \cos^2\phi'\}]^{\frac{1}{2}}}{4 - 2\eta(2 - \cos^2\phi') + \eta^2(1 - \cos^2\phi')}, \quad (89)$$

so that, writing y for $\cos\phi'$,

$$\frac{2(1-\mu'_0)}{(t-1)\eta} = \frac{2(t-1)\eta + y^2(2\eta - \eta^2) - (t-1)\eta^2 - y[(2\eta - \eta^2) \times \{4(t-1)\eta + \eta^2(1-t^2) + y^2(2\eta - \eta^2)\}]^{\frac{1}{2}}}{2(t-1)\eta [1 - \eta(1 - y^2/2) + \eta^2(1 - y^2)/4]}.$$

This is still exact, but it can be written as

$$\frac{2(1-\mu'_0)}{(t-1)\eta} = [A(y^2) - yB(y^2)] \{1 + O(\eta)\}, \quad (90)$$

where
$$A(y^2) = 1 + \frac{y^2(1-\eta/2)}{(t-1)}, \tag{91}$$

$$B(y^2) = \left[\frac{y^2(1-\eta/2)^2}{(t-1)^2} + \frac{(2-\eta)}{(t-1)} \left(1 - \frac{\eta(1+t)}{4} \right) \right]^{\frac{1}{2}}. \tag{92}$$

Substituting from (90) into (88),

$$(1-\mu'_0)/|2-\eta-\mu'_0(2-\xi)| = \frac{(A-yB)\{1+O(\eta)\}}{2(1+A-yB)}. \tag{93}$$

This has to be integrated from $y = -1$ to $+1$ (cf. (87)). We add the integrands for those values of y which are opposite in sign but equal in magnitude. The combined integrand, to be integrated over positive values of y , is

$$\frac{\{A+A^2-y^2B^2\}}{(1+A)^2-y^2B^2} \{1+O(\eta)\} = \frac{1}{2} + O(\eta).$$

It only remains to consider the limits on y , set by the requirement that μ'_0 be real. From (89), the condition is

$$(2\eta-\eta^2)y^2+4(t-1)\eta+(1-t^2)\eta^2\geqslant 0.$$

For $t\geqslant 1$, all values of y from 0 to 1 are permitted. If

$$t\leqslant \frac{1}{2}(1-\eta/8-3\eta^2/32),$$

there are no possible values of y . For t between these limits, the minimum y allowed is $\{2(1-t)\}^{\frac{1}{2}}+O(\eta)$. Hence the second integral in (82), having been reduced to (87), becomes

$$\pi N^2 r_0^4 x e^{-N\sigma x} (1-2\gamma\nu\xi)f(t), \tag{94}$$

where
$$f(t) = \left. \begin{aligned} &0 \text{ for } t\leqslant \frac{1}{2}, \\ &= (2/\pi) \cos^{-1} [\{2(1-t)\}^{\frac{1}{2}}] \text{ for } \frac{1}{2}\leqslant t\leqslant 1, \\ &= 1 \text{ for } t\geqslant 1. \end{aligned} \right\} \tag{95}$$

It remains to evaluate the first integral in (82), namely,

$$Nr_0^2 \iint L(x,\mu',\xi-1+\mu_1) \exp\left[-\frac{2\nu\gamma(1-\mu_1)}{1+\gamma\xi-\gamma(1-\mu_1)}\right] d\mu' d\phi', \tag{96}$$

with
$$\mu_1 = \mu\mu' + \sin\theta \sin\theta' \cos\phi'. \tag{97}$$

The expression (96) is of order s and need be evaluated only to that order. The integration over μ' itself provides a factor of order s , so the integrand need be known only to terms of order s^0 . Using (84), we reduce (96) to

$$Nr_0^2 \int_0^2 \int_0^{2\pi} M(x,\xi'/\eta') d\eta' d\phi', \tag{98}$$

where $\eta' = 1-\mu'$ and $\xi' = \xi-1+\mu_1$. Writing also $\eta_1 = 1-\mu_1$ we obtain

$$\xi'/\eta' = (\xi-\eta)/\eta' - 1 - 2\cos\phi' (\eta/\eta')^{\frac{1}{2}} + O(\eta)$$

which reduces (98) to

$$Nr_0^2 \eta \int_{z=0}^{\infty} \int_0^{2\pi} M\left(x, \frac{t-1}{z} - 1 + \frac{2}{\sqrt{z}} \cos\phi'\right) dz d\phi', \tag{99}$$

where $z = \eta'/\eta$.

Substituting (85), (94) and (99) into (82), the terms of order s^0 give

$$\frac{\partial M}{\partial x} + N\sigma M = \pi N^2 r_0^4 x f(t) e^{-N\sigma x},$$

of which the solution with $M = 0$ at $x = 0$ is

$$M(x, t) = \frac{\pi}{2} (Nr_0^2 x)^2 f(t) e^{-N\sigma x}. \quad (100)$$

The terms in (82) of order s give a solution for $P(x, t)$, using the boundary condition that $P = 0$ at $x = 0$. The correct form is

$$P(x, t) = e^{-N\sigma x} [(Nr_0^2 x)^2 R(t) + (Nr_0^2 x)^3 S(t)], \quad (101)$$

with
$$R(t) = \frac{\pi f(t)}{2\sqrt{1+t^2}} (1 - 2\gamma vt) \quad (102)$$

and

$$S(t) = \frac{\pi}{6\sqrt{1+t^2}} \int_0^\infty \int_0^{2\pi} f\left\{\frac{t-1}{z} - 1 + \frac{2}{\sqrt{z}} \cos \phi'\right\} dz d\phi' - \frac{\pi f(t)}{6\sqrt{1+t^2}} \left[\frac{\sigma}{r_0^2} + \frac{t}{r_0^2} \left\{ \frac{\partial \sigma}{\partial \xi} \right\}_0 \right]. \quad (103)$$

This completes the solution for small deflexions and energy degradations.

E3. The relative dose

The relative dose is

$$\text{Relative dose}(x) = \int_0^\gamma \int_{-1}^1 (E/\gamma) \sigma_a(E) I(x, \mu, E) dE d\mu / \sigma_a(\gamma),$$

giving $e^{-N\sigma x}$ from unscattered quanta, while the contribution from radiation scattered once is (29) of § B 2. From all higher orders we get

$$\int_0^\gamma \int_{-1}^1 (E/\gamma) \sigma_a(E) J(x, \mu, E) dE d\mu / \sigma_a(\gamma) = \int_0^\infty \int_{-1}^1 (E/\gamma) \sigma_a(E) L(x, \mu, \xi) d\xi d\mu / \sigma_a(\gamma). \quad (104)$$

We write

$$L(x, \mu, \xi) = (\pi/2) Y^2 e^{-D} f(t) \exp \left[-\frac{\sigma \eta Y}{3r_0^2} - \xi \left\{ 2\gamma v + \frac{Y}{3r_0^2} \left\{ \frac{\partial \sigma}{\partial \xi} \right\}_0 \right\} \right] \\ + \frac{s\pi Y^3 e^{-D}}{6\sqrt{1+t^2}} \int_{z=0}^\infty \int_0^{2\pi} f\left\{\frac{t-1}{z} - 1 + \frac{2}{\sqrt{z}} \cos \phi'\right\} dz d\phi', \quad (105)$$

where $Y = Nr_0^2 x$ and $D = N\sigma x$. This expression for L is correct to terms of order ξ and η and is always positive or zero, which is a necessary property of the intensity.

We have, approximately,

$$E\sigma_a(E)/\gamma\sigma_a(\gamma) = e^{-G\xi}, \quad (106)$$

where

$$G = - \left\{ \frac{\partial}{\partial \xi} \ln (E\sigma_a) \right\}_{\xi=0} > 0.$$

This is correct to first order in ξ and quite a fair approximation over a wide range of E ; for example, if $\gamma = 10$, $G = 4.62$, and (106) gives 0.50 instead of 0.60 at the value $E = 4$, where $2\gamma\xi$ has already attained 3 and the intensity (105) is itself far from rigorous.

Substituting (105) and (106) in the relative dose (104), two integrals appear which are simplified by writing $\xi = s \sin \lambda$, $\eta = s \cos \lambda$, $t = \tan \lambda$ and the relative dose becomes

$$(\pi/2) Y^2 e^{-D} \int_0^\infty f(t) dt \left/ \left\{ \frac{D}{3} - 1 + Ht \right\}^2 \right. \\ + \frac{\pi Y^3 e^{-D}}{3G^3} \int_0^\infty t^{-3} dt \int_{z=0}^\infty \int_0^{2\pi} f \left\{ \frac{t-1}{z} - 1 + \frac{2}{\sqrt{z}} \cos \phi' \right\} dz d\phi', \quad (107)$$

where
$$H = G + 2\gamma\nu + \frac{D}{3\sigma} \left(\frac{\partial \sigma}{\partial \xi} \right)_0. \quad (108)$$

For $\gamma = 10$, H is roughly $2.5 + 1.7D$.

(107) is our final expression for the dose from radiation scattered more than once, and is not difficult to compute. Actually our numerical results, quoted in §7, were computed using $(\frac{1}{3}D - Ht)^2$ in the denominator of the first integral in (107). This arose by approximating to $R(t)$ on the assumption $\gamma\nu \gg 1$, which is rather more stringent than the assumption $2\gamma\nu \gg 1$ made elsewhere in the derivation. One cannot expect the slight change made in (107) to do much to improve the agreement with experiment. The assumption $2\gamma\nu \gg 1$ is itself rather stringent, since $\nu \sim 0.15$ seems to be needed. This makes $\gamma = 5$ a lower limit for the derivation of (107).

REFERENCES

- Bethe, H. A., Fano, U. & Karr, P. R. 1949 *Phys. Rev.* **76**, 538.
 Bobrowsky, A. R. 1948 *Tech. Note Nat. Adv. Comm. Aero., Wash.*, no. 1712.
 Chandrasekhar, S. 1946a *Astrophys. J.* **104**, 110.
 Chandrasekhar, S. 1946b *Astrophys. J.* **104**, 191.
 Cowan, C. L. 1948 *Phys. Rev.* **74**, 1841.
 Ellis, F. 1946 *Brit. Med. Bull.* **4**, no. 1, BMB 804.
 Fowler, W. A., Lauritsen, C. C. & Lauritsen, T. 1948 *Rev. Mod. Phys.* **20**, 236.
 Heitler, W. 1944 *Quantum theory of radiation*, 2nd ed. pp. 194, 214 and 258.
 Hirschfelder, J. O. & Adams, E. N. 1948 *Phys. Rev.* **73**, 863.
 Hirschfelder, J. O., Magee, J. L. & Hull, M. H. 1948 *Phys. Rev.* **73**, 852.
 van der Hulst, H. C. 1948 *Astrophys. J.* **107**, 220.
 Kaye, G. W. C., Binks, W. & Bell, G. E. 1936 *Brit. J. Radiol.* **9**, 161.
 Klein, O. & Nishina, Y. 1929 *Z. Phys.* **52**, 853.
 Mayneord, W. V. 1940 *Brit. J. Radiol.* **13**, 235.
 Mayneord, W. V. & Clarkson, J. R. 1944a *Brit. J. Radiol.* **17**, 151.
 Mayneord, W. V. & Clarkson, J. R. 1944b *Brit. J. Radiol.* **17**, 177.
 Ornstein, L. S. 1937 *Mon. Not. R. Astr. Soc.* **97**, 207.
 Schuster, A. 1905 *Astrophys. J.* **21**, 1.
 Singer, G., Braestrup, C. B. & Wyckoff, H. O. 1946 *Bur. Stand. J. Res., Wash.*, **37**, 147.
 Wheeler, J. A. & Lamb, W. E. 1939 *Phys. Rev.* **55**, 858.
 Wightman, A. 1948 *Phys. Rev.* **74**, 1813.
 Wyckoff, H. O., Kennedy, R. J. & Bradford, W. R. 1948 *Nucleonics* (Nov.), p. 62.

The velocity of propagation of electromagnetic waves derived from the resonant frequencies of a cylindrical cavity resonator

By L. ESSEN, D.Sc.

The National Physical Laboratory

(Communicated by E. C. Bullard, F.R.S.—Received 24 June 1950)

[Plate 8]

The cavity resonator used in this investigation is a silver-plated steel cylinder 6.5 cm. in diameter and of adjustable length. Resonance in the H_{011} mode is established at a frequency in the region of 9000 Mc./sec., and the length is then varied to give successive resonances at half wave-length intervals. The wave-length is thus determined and this, together with the frequency, the diameter and a correction term involving the sharpness of resonance, enables the velocity to be calculated. This procedure has some advantage over that used previously by Essen & Gordon-Smith in which the measurements were made with a resonator of fixed dimensions. The wave-length is determined only from differences in length, the first resonant length not being used, and in this way certain end-effects, such as those due to the coupling loops and to surface imperfections, are eliminated or greatly reduced. Moreover, by using different frequencies, or different modes at the same frequency, the diameter can be eliminated from the calculations and a value of c thus obtained in terms of frequency and length both of which can be measured with high precision.

The result obtained is $299,792.5 \pm 3$ km./sec., and is thus in close agreement with that obtained by Essen & Gordon-Smith with a fixed cavity and also with the value of c determined recently by Bergstrand with an optical method.

1. INTRODUCTION

The value obtained previously by the author (Essen 1947; Essen & Gordon-Smith 1948) was $299,792 \pm 9$ km./sec., whereas the commonly accepted value for the velocity of light is $299,776 \pm 4$ km./sec. (Birge 1941). It was decided therefore to repeat the experiment using a rather different technique designed to reduce the limits of error. In the meantime the author's value has been supported by the results of an optical method (Bergstrand 1949*a, b*) giving $299,796 \pm 2$ km./sec. and a radar method (Aslakson 1949*a, b*) giving $299,792 \pm 2.4$ km./sec. On the other hand, in a recent paper on the values of fundamental constants, Du Mond & Cohen (1948) state that there is no need to re-examine Birge's value which is indeed supported by the preliminary results of a cavity resonator method due to W. W. Hansen at Stanford University. There is thus still a considerable amount of doubt about the true value of this important constant, and it is hoped that the work described in this paper will help to reduce it.*

The main interest in a paper of this kind is centred on the result and its probable accuracy, and the experimental procedure and observational details are therefore given more fully than might otherwise be justified.

* Since the completion of this work additional results of 299,793 km./sec. by Bergstrand (1950*a, b*) and 299,775 km./sec. by Houston (1949) have been announced. The result obtained in the present work was reported in *Nature* (Essen 1950).

2. GENERAL DESCRIPTION OF THE METHOD AND APPARATUS

The method is based on the electrical resonance of a length of cylindrical wave-guide closed by metal disks at both ends. The resonant frequencies depend on the internal dimensions and the free-space velocity c according to the relationship

$$c = \frac{f'_{lmn} \left(1 + \frac{1}{2Q}\right)}{\sqrt{\left[\left(\frac{\tau}{\pi D}\right)^2 + \left(\frac{n}{2L}\right)^2\right]}}, \quad (1)$$

where f'_{lmn} is the measured frequency of the mode of resonance denoted by the suffix lmn , D and L are the diameter and length of the resonator, Q is the sharpness of resonance ($\pi/\log \text{dec.}$) and τ is a constant for a series of particular lm modes. l , m , and n characterize the shape of the electric field, l and m being the number of periodic variations in the angular and radial directions, and n in the axial direction. That is, n is the number of half wave-lengths of the standing wave-field pattern in the length of the cylinder, and $2L/n$ is equal to the guide wave-length λ .

The experiment therefore consists in the measurement of the dimensions, resonant frequency, and Q .

To excite and observe the resonance it is necessary to introduce into the resonator small probes or loops coupled respectively to the source of oscillations and a detector. These alter slightly the geometry of the resonator and the configuration of the field, and therefore affect the resonant frequency. This effect has been eliminated in the present form of the apparatus shown in figure 1. The length of the resonator R is altered by the movement of a piston P , and successive resonances are obtained at values of L corresponding to values of n of 1, 2, 3, ... in equation (1). The length for the first resonance is affected by the coupling loops, and the gap between the piston and the walls of the cylinder and is not used in the measurement of λ and c . It has already been shown experimentally (Essen 1946) that even when the length of the first resonance is in error by 2 % the distance between successive resonances agrees with the theoretical value to within 0.01 %. In the conditions of the present experiment the length at the first resonance is in error by a far smaller amount, and it can be assumed that the value of λ determined by differences is not affected to an appreciable extent by the end conditions.

The most suitable modes of resonance for this work are the H_{0mn} modes, because the electric field for these is circular and concentric with the axis of the cylinder. There is no electric field across the gap between the piston and the walls and therefore no coupling with the coaxial line circuit below the piston. The effect of the gap on the resonant frequency is simply that due to the geometrical difference from a perfect cylinder, and as explained above is eliminated by using only the different lengths for successive resonances.

With the resonator described and the available oscillators the H_{01n} ($n = 1$ to 8) and H_{02n} ($n = 1$ to 3) modes could be used.

A useful feature of the apparatus is that it enables the diameter of the cylinder to be measured in terms of length and frequency. If λ is determined at two frequencies

for the same H_{0m} family of resonances (same value of τ) D can be calculated from the expression, obtained from (1),

$$D = \frac{\lambda_1 \lambda_2 \tau}{\pi} \left[\frac{f_2^2 - f_1^2}{(\lambda_1 f_1)^2 - (\lambda_2 f_2)^2} \right]^{\frac{1}{2}}. \tag{2}$$

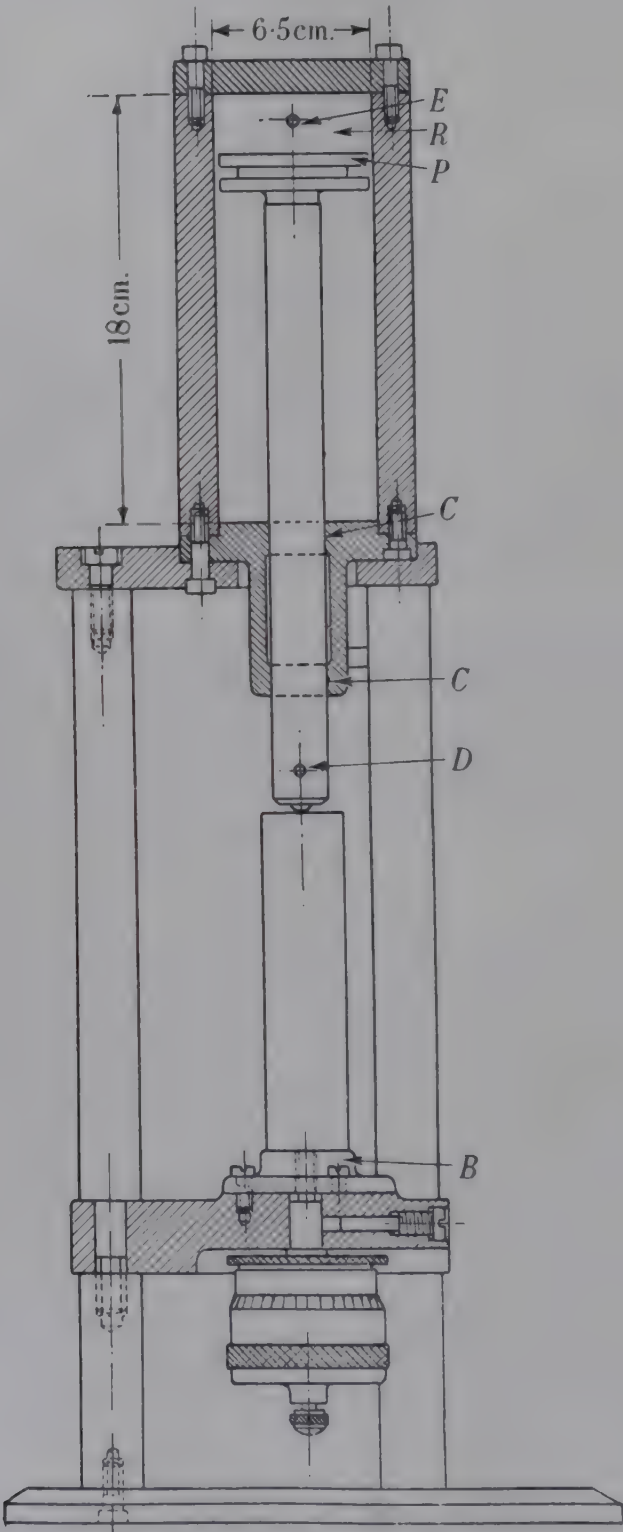


FIGURE 1.

The values of f here are the measured frequencies corrected for Q . Suitable frequencies for this method are 5960 and 9500 Mc./sec. Alternatively, the diameter can be measured in terms of length alone by observations on different modes of resonance

at the same frequency. We then have from (1) when the resonant frequencies corrected for Q are equal,

$$D = \frac{1}{\pi} \left[\frac{\tau_2^2 - \tau_1^2}{\left(\frac{1}{\lambda_3}\right)^2 - \left(\frac{1}{\lambda_4}\right)^2} \right]^{\frac{1}{2}}. \quad (3)$$

In practice frequency measurements are involved just as before in order to ensure that the values of λ refer to exactly the same frequency. The appropriate resonant conditions are obtained at a frequency of 10,830 Mc./sec. During the course of this work the accuracy of the metrological measurement was increased so that the final estimated accuracies are about the same in both cases. The electrical measurements, however, provide a useful check that the effective electrical diameter is the same as the mean metrological value.

The diameter is of course a constant for the experiments. The three variables are length, frequency and τ . The changes in length are measured by means of slip gauges between the ball-ended piston and the base-plate B . The lengths can thus be varied in steps of 0.0001 in. The precise condition of resonance is established by a slight adjustment of the frequency of the source to give a maximum indication in the detector, and the frequency is then measured in terms of the standard.

3. DETAILS OF THE RESONATOR

Figure 1 is a scale drawing of the cylindrical resonator and its mounting. The cylinder is made from case-hardened mild steel, ground and honed to be uniform in diameter to $\pm 0.25 \mu$ ($\pm 1 \times 10^{-5}$ in.). The inside surface is silver-plated to a thickness of about 5μ (2×10^{-4} in.) to give the necessary high electrical conductivity. Unfortunately, the plating disturbed the uniformity more than was expected, but the final variations in diameter, $\pm 1.6 \mu$ ($\pm 6 \times 10^{-5}$ in.), are still considerably less than those of the copper cylinder used in the previous experiment. The top disk, which is also lapped and silver-plated, is screwed to the cylinder through ample clearance holes to avoid any lateral strain and distortion. The steel piston P is lapped and plated, but the piston rod is left unplated so that the best possible sliding fit is obtained in the bearing C . The gap between the piston and the walls is 51μ (2×10^{-3} in.). The piston is shaped to reduce the coupling between the cylindrical resonator above it and the coaxial resonator below it. This is not necessary in the case of the H_0 modes but has been provided to make the apparatus suitable for investigating other modes of resonance for which the electric field crosses the gap.

The cylinder is supported on a heavy steel stand which also carries the lapped base-plate B , to which the slip gauges are wrung. The stem of a micrometer passes through a central hole in the base-plate and can be used for obtaining a continuous range of length alteration without using inconveniently small gauges. This facility was used only during preliminary explorations of the modes of resonance, and the stem was always withdrawn clear of the base-plate for the actual measurements. There is a bar D at the lower end of the piston rod to which are attached wires passing over pulleys. Counter-weights are adjusted so that the piston moves very slowly down when the slip gauges are removed. This facilitates the changing of the gauges

and ensures that a small uniform pressure is exerted on them. Four symmetrically placed coupling holes E of diameter 0.23 cm. (0.092 in.) are drilled round a circumference 1.1 cm. from the top of the cylinder. Coupling loops are inserted through two of them; the others are for alternative coupling positions and also to make sure that there is no appreciable difference in pressure between the inside and outside of the cylinder when this is placed in a chamber being evacuated. This chamber can be seen above the resonator in figure 3, plate 8 together with the mechanical arrangements for lowering and raising it.

4. MEASUREMENT OF RESONANT FREQUENCIES

The equipment shown to the left of figure 3 was developed for measuring any frequency in the range 300 to 25,000 Mc./sec.; and the relevant parts of it have proved very useful in the present work. Similar equipments have been described elsewhere (e.g. Husten & Lyons 1948) and no details need be given here. It generates frequencies at intervals of about 1 % throughout the range directly from the quartz frequency standard. The gaps between these harmonic frequencies are filled by including in the circuit an interpolation oscillator. The frequency of this is controlled by a dial the readings of which give the final frequency with an accuracy of 1 part in 10^8 . The equipment is shown as the standard frequency oscillator in figure 2*a*. It is coupled to a receiver (Spectrum analyzer, model TSX-4SE) in which the received signal is displayed as a vertical deflexion on a cathode-ray tube. As the frequency is varied this deflexion moves horizontally across the screen, its amplitude remaining constant.

An auxiliary oscillator (CV 129) is used to supply energy to the resonator to which it is coupled by a small loop. A similar loop couples the resonator to the receiver. There is a negligible direct coupling between these loops, and no signal from the resonator appears in the receiver until resonance is established. As the frequency is varied through the resonant value the deflexion passes through a maximum value, as well as moving slightly across the screen. The setting to a maximum is rather critical. If, for example, the Q of the resonator is 50,000, the deflexion decreases to one-half of its peak value for a frequency change of 1 part in 10^5 . To enable the setting to be made with the necessary precision the CV 129 oscillator is fitted with a fine frequency control driven by a micrometer head. When the frequency of the oscillator has been set to give a maximum deflexion the standard is varied until its deflexion coincides with that from the resonator. This setting can be made very precisely because the receiver is approximately a square-law instrument, and when the two signals are superimposed there is a large increase in the amplitude of the combined signal. It is estimated that with Q 's of the order of 50,000 the resonant frequency can be measured to a few parts in 10^7 .

In some of the measurements the standard-frequency source itself was used to energize the resonator, and in this case it was merely necessary to vary the frequency to give a maximum deflexion on the indicator and then read off the value from the dial; but in the 9000 Mc./sec. region the power available from the standard is small, and this method was in general less convenient than that using an auxiliary oscillator.

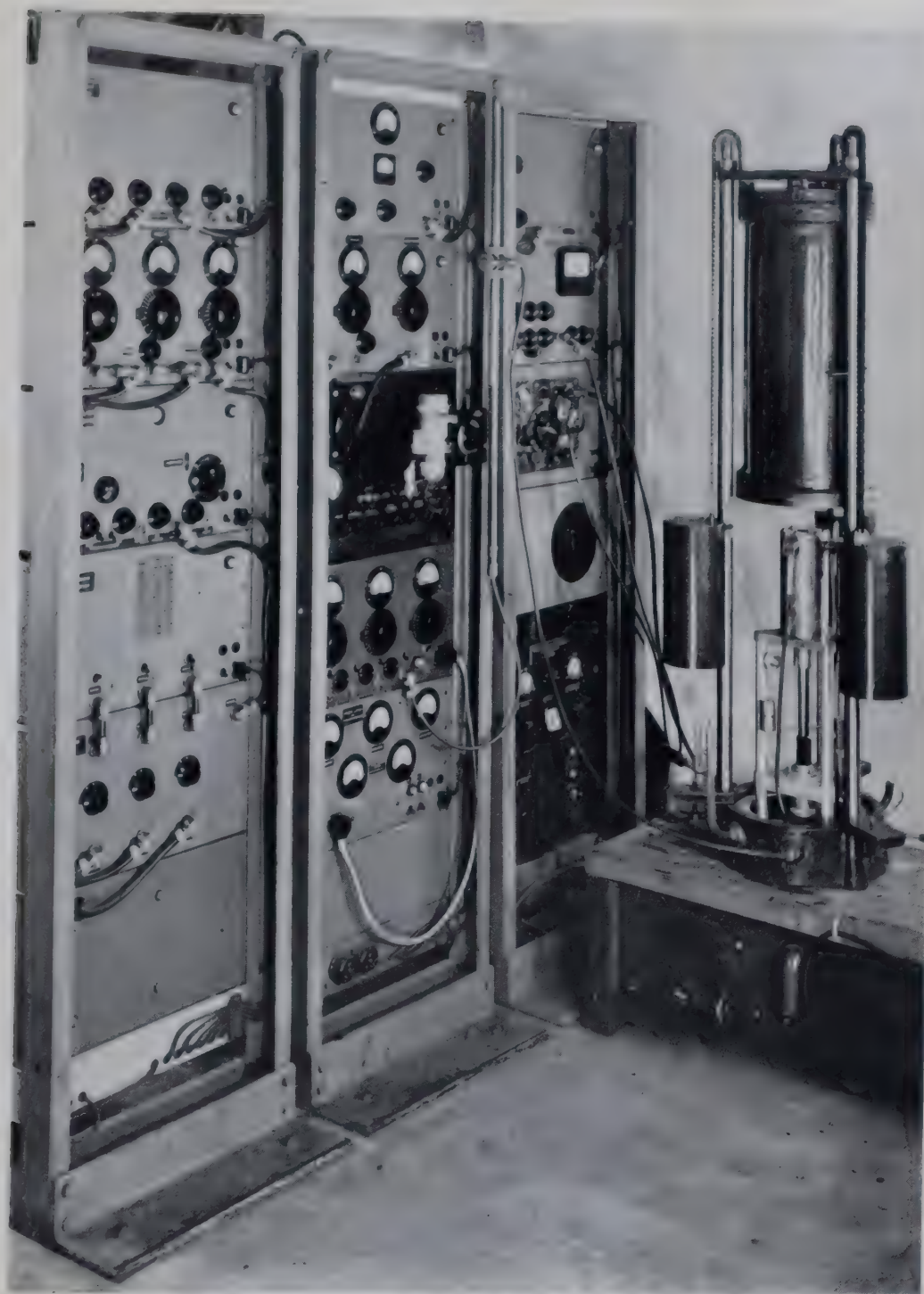


FIGURE 3. Cavity resonator and micro-wave frequency standard. The resonator is on the right, together with the cover which can be lowered and evacuated. The resonator is excited by a harmonic of a quartz-controlled oscillation multiplied up to the appropriate frequency in the equipment on the left, which also contains the receiver used for detecting the resonance.

No suitable receivers or oscillators were available for the measurements at 5960 Mc./sec., and a slightly different technique was used. The standard frequency was square-wave modulated at a frequency of 1000 c./sec. and used to energize the resonator directly. The output loop from the resonator was fed to a crystal diode detector, and the 1000 c./sec. signal obtained was amplified to give a deflexion on a cathode-ray tube. The maximum deflexion indicated the condition of resonance. The apparatus in figure 3 is arranged for this method, and no auxiliary oscillator is therefore shown.

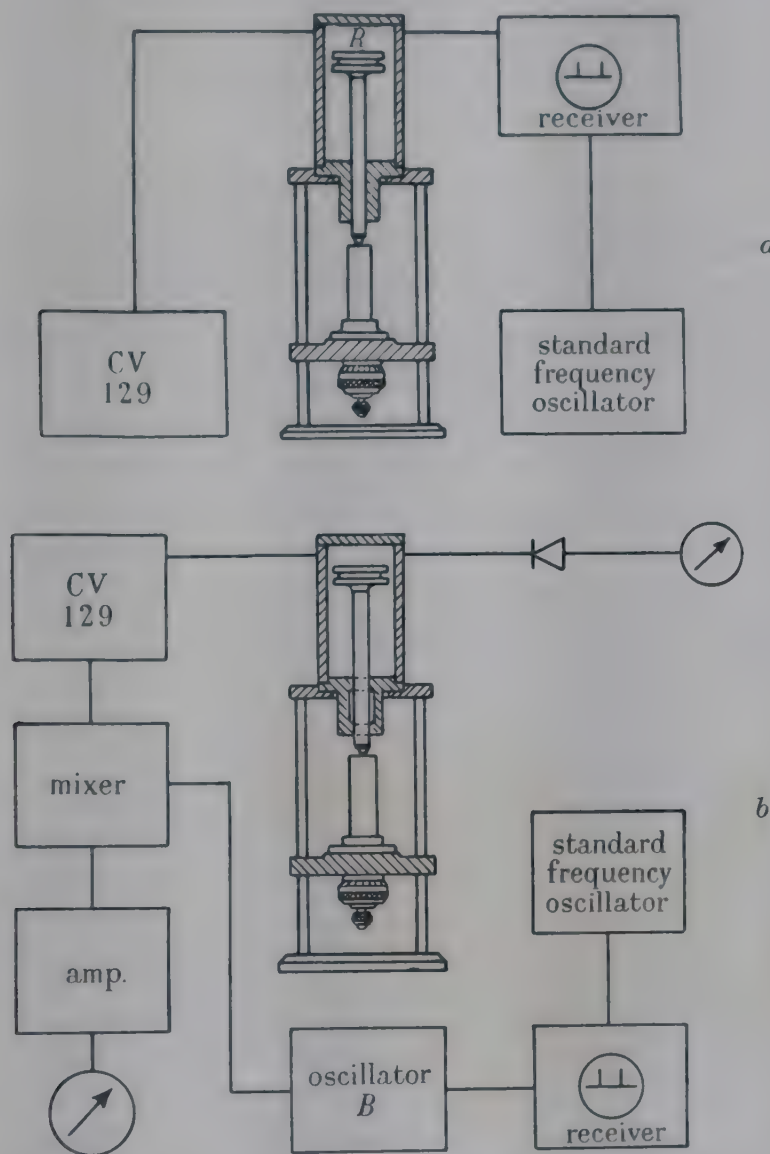


FIGURE 2. Equipment for measurement of resonant frequencies.

At 10,830 Mc./sec. none of these techniques could be used. A CV 129 valve oscillator was specially adjusted to operate in the required frequency range. It was coupled to the resonator as shown in figure 2b, and the output was detected by a diode to give a deflexion on a sensitive galvanometer. An additional oscillator *B* was set to give a zero heterodyne beat frequency between its fourth harmonic and the CV 129 oscillator. The frequency of *B* was then measured by the standard-frequency oscillator as before, a spectrum analyzer of suitable range being used as the receiver.

5. THE Q OF THE RESONATOR

The values of Q were measured with an arrangement similar to that just described and shown in figure 2*b*, because for this measurement it is necessary to know the law of the detector. It is known that the current given by a crystal diode is very closely proportional to the square of the input oscillatory voltage for the range of currents used and that the galvanometer deflexion is proportional to the current. The system therefore is a square-law detector, and points on the resonant curve where the deflexion is one-half of the peak deflexion are the half-power points. There was, however, no need to use the additional oscillator B and the associated mixer and amplifier when there was a receiver in the range of the resonant frequency. The two frequencies corresponding to the half-power points give the Q which is simply $f/(f_1 - f_2)$.

All the values of Q were found to be very nearly 75 % of their theoretical values, and they were therefore rounded to this value. It is of some importance to consider why the Q 's are less than their theoretical values, because if the reduction is due to a thin layer of higher resistivity on the surface of the walls of the resonator the correction for the field penetration will not necessarily be $(1 + 1/2Q)$ (Whinnery 1942). The silver plating on the cylinder used in the present experiment is about ten times the effective skin depth at the frequencies of measurement, and from the electrical point of view can be regarded as of infinite thickness. It was not at first anticipated that the surface layer formed on the silver on exposure to air would cause an appreciable error, but the results indicated that this assumption may be wrong, and the effect is discussed in more detail in § 12.

6. THE MEASUREMENT OF LENGTH

Throughout the experiment many combinations of slip gauges were used to set the position of the piston. It was not convenient to have these blocks of gauges calibrated as they were used, but the individual gauges were calibrated immediately after the electrical measurements by an interferometer method in the Metrology Division of the Laboratory. Tests were also made to check that the author's technique of wringing the gauges together did not introduce any additional error. If the gauges and the piston (figure 1) are not exactly in line the movement of the piston will be greater than the change in length of the gauges by a factor equal to the secant of the angle between them. The Metrology Division made a direct measurement of this effect. An interferometer method was devised which is likely to be of interest in the field of precision metrology and is therefore the subject of a separate note (Barrell & Puttock 1950). It was found that the actual movement of the piston was greater than the length of the gauges by a factor of $2.4 \times 10^{-6} \pm 1 \times 10^{-6}$. The lengths were all corrected for the gauge errors and for the above secant effect.

7. UNIFORMITY OF THE GAP BETWEEN THE PISTON AND THE WALL

For the measurements of the movement of the piston described above the top disk was removed from the cylinder and interference fringes were observed in a polished plate fixed to the top of the piston. Any tilt of the piston during its

movements therefore resulted in a change in the number of fringes across the polished plate. It was deduced from the fringe movements that the tilt resulting from a definite sideways force applied to the rod was about 1.2×10^{-4} radian, but that if the piston was allowed to fall freely the variation of tilt was only of the order of 5×10^{-6} radian. If it was raised above the height required for a particular setting and allowed to fall slowly on to the gauges the tilt was of this order. This corresponds to a lateral movement of the piston when in the worst position near the top of the cylinder of 2.5×10^{-5} in. A direct check during the electrical measurements showed that, as would be expected from the above results, a sideways pressure on the piston produced a negligible effect on resonant frequency.

8. MEASUREMENT OF DIAMETER

The diameter was measured in the Metrology Division of the Laboratory by a method based on that of Dymott & Jennings (1948). The apparatus needed considerable modification to deal with a cylinder as large as that used in this experiment and to give the required accuracy. The instrument finally developed will be described elsewhere in a separate note. The diameter was measured in seventy-eight positions and the results are given in table 1.

TABLE 1. INTERNAL DIAMETER OF SILVER-PLATED STEEL CYLINDER

axial position of measurement (distance from end containing coupling holes) (cm.)	measured diameter at 20° C in six symmetrically disposed diametral planes (unit = 1 cm.)					
	1	2	3	4	5	6
0.3	6.51766	6.51756	6.51748	6.51748	6.51733	6.51759
1.6	46	51	48	45	43	42
2.8	56	58	50	42	51	58
4.1	56	57	50	44	49	56
5.4	56	58	48	46	50	46
6.7	59	56	44	47	43	59
7.9	51	47	42	49	43	52
9.2	54	49	47	48	40	52
10.5	56	41	40	47	46	43
12.1	58	55	44	46	52	49
13.0	62	54	62	62	60	48
14.0	59	53	65	66	63	54
15.6	53	47	50	57	52	46

The mean value is 6.51752 cm. and the extreme errors are considered to be ± 0.00003 cm.

9. EXPERIMENTAL PROCEDURE

The measurements were carried out in a temperature-controlled room, but the apparatus itself was not separately controlled. Thermocouples were fixed at three positions on the cylinder and its supports to check the temperature gradients in the

system and to give its mean temperature. When the gauges had been wrung in position the chamber was lowered and evacuated, and several hours were allowed for steady conditions to be reached before any frequency measurements were made.

The aim was to obtain a series of resonances such as the $H_{011}, H_{012}, \dots, H_{01n}$ at a fixed frequency and so obtain the wave-length corresponding to that frequency; but in practice it would be too tedious to adjust the gauges to give the resonant condition. The approximate gauge lengths required were therefore determined by calculation and the resonant condition established by a slight frequency adjustment. The experiment thus yielded n different resonant frequencies corresponding to the n gauge lengths. Each measurement was probably at a slightly different temperature, and the values of Q were also different for each mode. The results were adjusted in the following way:

(1) The frequencies were corrected for the temperature deviation from 20°C . The coefficient measured over a range of 5°C was -1 part in 10^5 per 1°C rise, and was the same within the accuracy of the measurements at a number of positions of the piston. If the whole apparatus were made from the same material the temperature coefficient of frequency would have the same value, with sign reversed, as the linear expansion of the metal; but a direct check was made because of the different types of steel used in the construction.

(2) The correction $f/2Q$ was applied. We then have the resonant frequencies corresponding to various lengths of a perfect cylinder.

(3) The lengths already corrected for gauge errors and the secant error mentioned in §6 were adjusted to correspond to one frequency near the average of the measured values. The law for this adjustment obtained from (1) is

$$\frac{\delta L}{\delta f} = \frac{nf}{2c^2} \left[\left(\frac{f}{c} \right)^2 - \left(\frac{\tau}{\pi D} \right)^2 \right]^{-\frac{1}{2}} \text{ cm./c./sec.}, \quad (4)$$

$1/2Q$ being negligible for the purpose of this adjustment. The numerical values are not required with great accuracy as the amount of the adjustment is only a few parts in 10^5 . The length measurements in the tables are expressed throughout in inches because it was then clear which gauges were used and easy to apply their appropriate corrections.

A typical set of results is given in table 2 in which the above procedure can be followed. It will be seen that the successive half-wave lengths show no systematic deviations as the length increases, although the scatter of the order of $\pm 2 \times 10^{-5}$ in. is rather larger than expected. A part of this is due to the variations of diameter along the cylinder, and it also includes the errors due to temperature, frequency measurement and the setting to resonance. If the total length of about 5 in. is in error by 2×10^{-5} in., the proportional error in c will be 3×10^{-6} . This includes all observational errors, excepting that of the diameter measurement. The mean value of the half wave-length is taken as the value of L , for $n = 1$, in expression (1). The mean diameter for the part of the cylinder used is the measured value found from table 1, and using the other relevant data in table 5, c is calculated.

This procedure was repeated at several different frequencies.

10. EXPERIMENTAL RESULTS

A good many preliminary measurements were made in air to gain experience of the technique and to find the most favourable conditions. When the results of these measurements are corrected for the permittivity of the air they are all in agreement with the later results *in vacuo* within ± 1.5 km./sec., but they have not been used in determining the result. After preliminary tests *in vacuo* several sets of observations were taken with particular care and the final result is based on these measurements. Table 2 gives the detailed observations at 9500 Mc./sec. made before and after the measurement of diameter. When the apparatus was dismantled for the diameter measurement the end-faces of the resonator were lapped as they were found to be slightly concave. The required gauge lengths in the two sets of observations are slightly different in consequence, but the velocity is the same within the limits of experimental accuracy. Table 3 gives similar results at 9000 Mc./sec. It was thought useful to take these measurements as a check because they involve different gauges and different settings of the frequency-measuring equipment. The results are seen to be in close agreement with those obtained at 9500 Mc./sec.

Measurements were then made at 5960 Mc./sec. At this frequency the wave-length in the guide is much longer, so that only the first two resonances could be obtained. It is the lowest frequency that could conveniently be used and is chosen to make the velocity depend as far as possible on diameter and not on the length measurements. Since only two resonances could be observed repeat measurements were made and the average value of L determined. As before, a second set of observations was taken after the metrological measurements. In this case the frequency also was altered slightly to remove the required resonances farther away from neighbouring modes. The two results are in good agreement, although they differ appreciably from those at the higher frequencies. To check whether this discrepancy was associated with the different frequency, measurements were made at 10,830 Mc./sec. At this frequency it is possible to observe the H_{02} as well as the H_{01} resonances. The H_{02} mode is rather similar to the H_{01} mode at 5960 Mc./sec. in that the half wave-length is large and c is determined mainly by the diameter. Unfortunately, the observations at 10,830 Mc./sec. were not so accurate as at the other frequencies because of the different technique that had to be employed and a scatter of ± 2 km./sec. was obtained. Detailed results of these measurements are not therefore given, but the averages of a number of measurements are included in the summarized results in column 4 of table 6. The same effect is clearly evident; a lower result is obtained when the diameter is the preponderant quantity in determining c . The ratios of the two parts of the denominator of expression (1) are shown in column 3 of this table. The author was fairly confident from the results of the preliminary tests with the apparatus that this discrepancy was not due to accidental experimental causes. However, a number of experiments were made to establish this more thoroughly.

11. FURTHER EXPERIMENTAL CHECKS

The effect of neighbouring modes had not been examined deliberately, although it would almost certainly have shown up in the course of the work. The conditions

TABLE 2. RESULTS FOR H_{01n} MODE AT 9500 Mc./SEC.

date	mode	temp. (° C)	gauge length nominal (in.)	gauge length corrected (in.)	frequency measured (Mc./sec.)	frequency adjusted to 20° C	$F/2Q$ (Mc./sec.)	frequency for 20° C and perfect resonator	length adjusted for 9498.3 Mc./sec. (in.)	length difference = $\lambda/2$ (in.)
31. v. 49	H_{018}	18.7	0.3926	0.392602	9498.039	9497.915	0.076	9497.991	0.392910	—
31. v. 49	H_{017}	19.0	1.1626	1.162594	9497.992	9497.897	0.085	9497.982	1.162876	0.769966
31. v. 49	H_{016}	19.4	1.9326	1.932606	9497.954	9497.897	0.095	9497.992	1.932838	0.769962
31. v. 49	H_{015}	20.3	2.7026	2.702610	9497.777	9497.805	0.105	9497.910	2.702853	0.770015
31. v. 49	H_{014}	19.4	3.4726	3.472598	9497.819	9497.762	0.114	9497.876	3.472812	0.769962
1. vi. 49	H_{013}	19.1	4.2426	4.242600	9497.859	9497.774	0.133	9497.907	4.242747	0.769935
1. vi. 49	H_{012}	19.5	5.0126	5.012588	9497.610	9497.563	0.181	9497.744	5.012727	0.769980
2. vi. 49	H_{011}	18.6	5.7826	5.782592	9497.327	9497.194	0.322	9497.516	5.782690	0.769963
average value of $\lambda/2 = 0.769969$ in., $c = 299,789.5$ km./sec.										
apparatus dismantled, faces trued, reassembled										
9. xi. 49	H_{018}	19.5	0.3932	0.393194	9498.219	9498.171	0.076	9498.247	0.393247	—
9. xi. 49	H_{017}	21.0	1.1632	1.163191	9498.120	9498.215	0.085	9498.300	1.163195	0.769948
9. xi. 49	H_{016}	20.9	1.9332	1.933198	9498.157	9498.242	0.095	9498.337	1.933170	0.769975
9. xi. 49	H_{015}	20.8	2.7032	2.703203	9498.219	9498.295	0.105	9498.400	2.703141	0.769971
8. xi. 49	H_{014}	21.8	3.4732	3.473195	9498.136	9498.305	0.114	9498.421	3.473134	0.769993
8. xi. 49	H_{013}	20.9	4.2432	4.243196	9498.345	9498.430	0.133	9498.563	4.243098	0.769964
8. xi. 49	H_{012}	20.4	5.0132	5.013181	9498.544	9498.582	0.181	9498.763	5.013065	0.769967
7. xi. 49	H_{011}	20.7	5.7832	5.783183	9498.912	9498.979	0.322	9499.301	5.783058	0.769993
average value of $\lambda/2 = 0.769973$ in. $c = 299,790.6$ km./sec.										

TABLE 3. RESULTS FOR H_{01n} MODE AT 9000 Mc./SEC.

date	mode	temp. (° C)	gauge length nominal (in.)	gauge length corrected (in.)	frequency measured (Mc./sec.)	frequency adjusted to 20° C	$F/2Q$ (Mc./sec.)	frequency for 20° C and perfect resonator	length adjusted for 8990.8 Mc./sec. (in.)	length difference = $\lambda/2$ (in.)
5. x. 49	H_{016}	21.5	1.1352	1.513198	8990.532	8990.667	0.082	8990.749	1.513245	—
5. x. 49	H_{015}	23.1	2.3532	2.353194	8990.408	8990.686	0.090	8990.776	2.353212	0.839967
6. x. 49	H_{014}	20.8	3.1932	3.193202	8990.631	8990.703	0.108	8990.811	3.193195	0.839983
6. x. 49	H_{013}	20.8	4.0332	4.033187	8990.615	8990.687	0.126	8990.813	4.033185	0.839990
7. x. 49	H_{012}	20.8	4.8732	4.873175	8990.663	8990.735	0.162	8990.897	4.873146	0.839961
7. x. 49	H_{011}	21.2	5.7132	5.713194	8990.958	8991.066	0.280	8991.346	5.713109	0.839963
average value of $\lambda/2 = 0.839973$ in., $c = 299790.0$ km./sec.										
apparatus dismantled and reassembled										
2. xi. 49	H_{017}	19.1	0.6732	0.673202	8990.782	8990.702	0.082	8990.784	0.673221	—
2. xi. 49	H_{016}	20.9	1.5132	1.513198	8990.637	8990.717	0.082	8990.799	1.513192	0.839971
2. xi. 49	H_{015}	20.4	2.3532	2.353194	8990.737	8990.773	0.090	8990.863	2.353146	0.839954
3. xi. 49	H_{014}	20.9	3.1932	3.193202	8990.723	8990.804	0.108	8990.912	3.193134	0.839988
3. xi. 49	H_{013}	21.5	4.0332	4.033187	8990.681	8990.816	0.126	8990.944	4.033121	0.839987
4. xi. 49	H_{012}	21.1	4.8732	4.873175	8990.865	8990.964	0.162	8991.126	4.873076	0.839955
4. xi. 49	H_{011}	21.8	5.7132	5.713194	8991.492	8991.653	0.280	8991.933	5.713021	0.839945

average value of $\lambda/2 = 0.839967$ in., $c = 299788.5$ km./sec.

TABLE 4. RESULTS FOR H_{012} MODE AT 5960 AND 5970 Mc./SEC.

TABLE 4. RESEMBLED AND DISMANTLED APPARATUS										
date	mode	temp. (°C)	gauge length nominal (in.)	gauge length corrected (in.)	frequency measured (Mc./sec.)	frequency adjusted to 20° C	$F/2Q$ (Mc./sec.)	frequency for 20° C and perfect resonator	length adjusted for 5959 Mc./sec. (in.)	length difference = $\lambda/2$ (in.)
3. vi. 49	H_{011}	18.45	3.616	3.615999	5959.158	5959.064	0.101	5959.155	3.615325	2.937196
8. vi. 49	H_{011}	20.9	—	—	5958.990	5959.043	—	—	—	—
8. vi. 49	H_{012}	20.2	0.678	0.677993	5958.878	5958.890	0.095	5958.983	0.678129	—
9. vi. 49	H_{012}	19.4	—	—	5958.922	5958.886	—	—	—	—
					value of $\lambda/2 = 2.937196$ in.,		c = 299,784.5 km./sec.			
apparatus dismantled, faces trued, reassembled										
14. x. 49	H_{011}	21.2	3.662	3.661992	5969.732	5969.803	0.101	5969.900	3.661575	2.891435
15. x. 49	H_{011}	22.0	—	—	5969.674	5969.794	—	—	—	—
14. x. 49	H_{012}	20.7	0.770	0.769970	5969.641	5969.683	0.095	5969.780	0.770140	—
15. x. 49	H_{012}	20.4	—	—	5969.662	5969.686	—	—	—	—
					value of $\lambda/2 = 2.891435$ in.,		c = 299,784.6 km./sec.			

had been selected so that other modes were as far removed as possible from the required ones, the nearest being at a frequency differing by more than 2 parts in 10^3 from the resonant frequency, and the majority being much further removed. To check the extent of any interaction between modes experiments were made in the region where the frequencies of two modes cross over. It was found that the resonant frequencies departed from a smooth curve by 1 part in 10^5 when they approached to within 1 part in 10^4 of each other. In the conditions of the experiment therefore the effect of neighbouring resonances must have been quite negligible. There is one mode, however, the E_{1mn} , which has the same frequency as the H_{0mn} mode, and the coupling had therefore been arranged so that no E modes should be excited. The E_{1mn} modes would be masked by the H_{0mn} modes, but other E modes not masked in this way were not detected as the length of the resonator was varied. Other observers have found that even when the E modes are not directly excited the two modes can interact producing double peaks and low Q 's, one cause of such interaction being a tilt of the end-faces of the cylinder. No evidence of such interaction was observed, but, as a precaution, these faces were lapped when the apparatus was dismantled for the metrological measurements. They were found to be concave to the extent of nearly 0.003 cm., out of square with the axis by about 0.002 cm. over the 6.5 cm. diameter, and the piston was also slightly out of round. All these deviations were reduced by a factor of 10; and it is seen from tables 2, 3 and 4 that this produced no significant effect on the results.

TABLE 5. VALUES OF RELEVANT CONSTANTS

τ for H_{01n} modes $J_1(x)=0$ 1st zero	3.831706
τ for H_{02n} modes $J_1(x)=0$ 2nd zero	7.015587
π	3.141593
1 cm.	0.39370147 in.
average diameter D	6.51752 cm.
$\delta L/\delta F$ at 9498 Mc./sec.	0.000125 in./Mc./sec.
8990	0.000153
5960	0.00435
5970	0.00418
10,830 H_{01}	0.000080
10,830 H_{02}	0.00151

It was also found that the resonant frequency of an H_{013} mode as it crossed the E_{212} mode showed no deviations from a smooth curve, within the experimental accuracy; and that the resonant frequency of the H_{013} mode did not change by more than 1 part in 10^6 when the coupling loops were rotated so as deliberately to excite the corresponding E_{113} mode. It seemed safe to conclude therefore that the degenerate E modes were not affecting the results.

The Metrology Division assisted in this overall check of the experiment by re-examining the provisional equipment set up for the measurement of diameter, and an unsuspected source of error was revealed. The diameter originally given was too small by 1μ (4×10^{-5} in.), but correcting for this error reduced the discrepancy in the results by only one-third. This correction has already been made in the diameters given in table 1 and the results given in tables 2 to 6, and is mentioned here

only as an illustration of how a measurement in radio engineering was the means of increasing the accuracy of an instrument designed for precision metrology. These various checks established that the results obtained with the different modes of resonance differ by an amount that cannot readily be explained by observational errors.

TABLE 6. SUMMARY OF RESULTS

1	2	3	4	5	6
frequency (Mc./sec.)	resonant modes	$(\tau/\pi D)^2$ $(1/2L)^2$	c	c using calculated diameter $= 6.51773 \text{ cm.}$	c using correction term $(1 + 2.8/2Q)$
9,500	$H_{011}-H_{018}$	0.54	299,790.0	299,793.3	299,792.3
9,000	$H_{011}-H_{017}$	0.64	299,789.3	299,792.9	299,791.5
5,960	$H_{011}-H_{012}$	7.8	299,784.6	299,793.1	299,792.6
10,830	$H_{021}-H_{023}$	8.9	299,785	299,794	299,792
	$H_{011}-H_{019}$	0.37	299,789	299,792	299,791

Mean value ignoring results at 10,830.0 Mc./sec., 299,792.5 km. sec. Estimated maximum error $\pm 3 \text{ km./sec.}$

12. AN EXPLANATION OF THE DISCREPANCY AND THE CORRECTED RESULT

It will be remembered that the experiment consists essentially in the measurement of the wave-length in the guide by the movement of a piston. Imperfections in the shape, or the state of the surface, of the top end-plate and of the piston will affect the resonant positions but not the wave-length. This is very easily checked experimentally by, for example, putting a disk of insulator on the piston. Although the measured lengths are slightly adjusted to reduce the observations to standard conditions in accordance with the calculated law of frequency change with length, this does not affect the validity of the argument, and the values of λ used in the calculations will therefore be taken as correct within the limits of experimental error. The error in the frequency measurement is also small and easily determined, and we are therefore left to consider the possibility of errors in the values of D and of the correction factor $(1 + 1/2Q)$. The correction factor allows for the penetration of the current into the walls of the resonator, and is based on the assumption that the material is of uniform resistivity. It is clear that if the walls are coated with a poor conductor the penetration will be greater, thus increasing the surface reactance; but that since the greater part of the current is still carried by the good conductor beyond the surface layer, the resistance and therefore $1/Q$ will not be increased in the same proportion. Under such conditions the correction factor will be greater than $1 + 1/2Q$. At a fixed frequency the penetration will be constant and can be regarded simply as an increase in the effective diameter of the resonator. Formula (1) can then be rewritten

$$c = \frac{f'_{lmn}}{\sqrt{\left[\left(\frac{\tau}{\pi(D+\delta D)}\right)^2 + \left(\frac{n}{2L}\right)^2\right]}}$$

(5)

the value of L obtained by differences being unaffected by surface penetration as explained above. δD represents the total increase in the effective diameter and includes the normal surface penetration, the additional penetration caused by

a surface layer of insulating or poor conductivity material, and any effects due to mechanical imperfections and non-uniformity. Now by making measurements with two different modes of resonance it is possible to measure the effective diameter as explained in § 2 and thus to obtain a value of c in terms of the length and frequency measurements alone, without using the metrological value of diameter or making any assumptions about the surface layers on the walls of the resonator. Such measurements were made at 10,830 Mc./sec. using the H_{01n} and H_{02n} modes, but unfortunately these results were not considered accurate enough for the purpose of determining a small correction term. It was decided to use instead the two sets of results at 5960 and 9000 Mc./sec., and at 5960 and 9500 Mc./sec., in view of their higher precision, and to bear in mind that the value of δD might then be in error by about 15 % if it varies with frequency. The normal skin effect contribution to δD which is known to vary as $1/\sqrt{f}$ can be taken care of by reinserting the correction term $(1 + 1/2Q)$ making (5)

$$c = \frac{f'_{lmn} \left(1 + \frac{1}{2Q}\right)}{\sqrt{\left[\left(\frac{\tau}{\pi(D + \delta D')}\right)^2 + \left(\frac{n}{2L}\right)^2\right]}}. \quad (6)$$

If $\delta D'$ does not vary with frequency—if, for example, it consists of a layer of dielectric—this formula is still exact in spite of the different frequencies used. From equations (6) and (2) and the values of $f'_{lmn}(1 + 1/2Q)$ and λ in tables 2 and 4 the value for $D + \delta D'$ is found to be 6.51774 cm., and from the values in tables 3 and 4, 6.51772 cm. The average of these measured values is thus 2.1×10^{-4} cm. greater than the metrological value. The values of c based on a diameter of 6.51773 cm. are given in column 5 of table 6. They may be regarded as extreme values, the other extremes being obtained by attributing the whole of the discrepancy to a conducting surface layer of higher resistivity than the bulk metal. $\delta D'$ will then vary with frequency in a similar way to the resistance and is best allowed for by increasing the factor $(1 + 1/2Q)$ to a value which will bring the results at different frequencies into agreement. The value is found to be $(1 + 2.8/2Q)$ and the results obtained using this factor and the metrological value of D are given in column 6 of table 6. It will be seen that the two sets of results differ on the average by only 1.2 km./sec.

It is not easy to decide which is the more probable value. An inspection of the surface revealed a faint yellow discoloration together with greyish patches. Now in the case of silver it is generally believed that the surface tarnish is mainly silver sulphide which is a very poor conductor (Bädeker 1907). The film is detected by a yellowish interference colour when it is about 0.03μ thick, but the action is progressive and the grey colour suggests the presence of a considerably thicker film. It was also noticed that the surface was not a perfect polish but possessed a marked granular structure. There is thus visual evidence of a thin film and also of mechanical imperfections which would tend to make the effective diameter larger than the measured value.

The fact that the measured Q 's are lower than the theoretical values indicates that the resistance of the walls is higher than that of bulk silver. Some recent work has provided evidence of the existence of films 0.2μ thick having a resistivity ten times

as great as that of the bulk metal (Chambers 1950), and a film of this nature would increase the surface reactance of the walls and the correction to be applied for this effect (Whinnery 1942).

It may be necessary to consider also the effect of the small irregularities in the diameter of the cylinder, which is in some places oval to the extent of nearly 2μ . This will clearly disturb the electric field which is circular in the H_0 modes used, and the velocity of these waves is therefore likely to be affected more by these dimensional irregularities than that of the E_0 waves used in the previous investigation. It seems reasonable to assume that the effect will be greater when the velocity is determined more by the diameter than the length, and that it will therefore be corrected by the procedure described of deriving the effective diameter from the electrical measurements.

It appears, therefore, that there are a number of small effects, none of them large enough to account for the discrepancy alone, but they all act in the same direction, and together they are sufficient to account for an effective diameter about 2μ larger than the measured metrological value.

Since there is evidence of some effects which are dependent on frequency and of some which are not, it is reasonable to believe that the true result lies between those in columns 5 and 6 of table 6, and the mean of these values is therefore taken as the final value of this investigation, the results obtained at 10,830 Mc./sec. being ignored because of their lower accuracy, although they provide useful evidence for the cause of the discrepancy and the validity of its correction.

The maximum error of the measurements of frequency and length combined, including uncertainties due to temperature, was found from the experiments to be $\pm 4 \times 10^{-6}$. The maximum error in the measurement of diameter by either the metrological or electrical method is estimated at $\pm 5 \times 10^{-6}$, giving an error in the value of c of $\pm 2 \times 10^{-6}$. An additional uncertainty of $\pm 2 \times 10^{-6}$, which is one-half of the difference between the results in columns 5 and 6 of table 6, will be allowed for residual errors caused by surface tarnish and other imperfections in the cylindrical walls of the resonator, making the estimated maximum error in the determination of $c \pm 8 \times 10^{-6}$ or approximately ± 3 km./sec.

Further experiments could be designed to examine the effects of surface imperfections. A slightly larger diameter, for example, would enable all the measurements to be made at a single frequency in the range of available oscillators and receivers; but, on the other hand, it is doubtful whether the cylinder could be made as uniform in dimensions as the present one, and whether therefore there would be any overall increase in accuracy.

On the whole it seems that further work along these lines is not justified at present, particularly as there is a possibility that other radio methods may give higher accuracies. For example, Barrell (1947) has suggested the use of radio interferometry for the measurement of length. If this technique can be developed to give the necessary precision of measurement for metrology work it will automatically give the value of velocity to about 1×10^{-6} . This follows from the fact that the frequency of the waves as well as their wave-length can be measured, which is not the case in optical interferometry.

The author desires to acknowledge the assistance of the Metrology workshop, and in particular of Mr White who made the resonator with extraordinary precision, and of Mr Barrell, in the Metrology Division, for undertaking various check measurements on the dimensions. He would also like to acknowledge the assistance of Mr E. G. Hope, in the Electricity Division, in observational work and of Dr L. Hartshorn for useful discussions on the presentation of the paper. The work described has been carried out as part of the Research Programme of the National Physical Laboratory and the paper is published by permission of the Director of the Laboratory.

REFERENCES

- Aslakson, C. I. 1949*a* *Nature*, **164**, 711.
Aslakson, C. I. 1949*b* *Trans. Amer. Geophys. Un.* **30**, 475.
Bädeker, K. 1907 *Ann. Phys., Lpz.*, **12**, 749.
Barrell, H. 1947 *Trans. Instrum. Meas. Conf., Stockholm*.
Barrell, H. & Puttock, M. J. 1950 *Brit. J. Appl. Phys.* **1**, 87-91.
Bergstrand, E. 1949*a* *Nature*, **163**, 338.
Bergstrand, E. 1949*b* *Ark. Mat. Astr. Fys. A*, **36**, 1.
Bergstrand, E. 1950*a* *Ark. Fys.* **2**, no. 15, pp. 119-150.
Bergstrand, E. 1950*b* *Nature*, **165**, 405.
Birge, R. T. 1941 *Rep. Phys. Soc. Progr. Phys.* **8**, 90.
Chambers, R. G. 1950 *Nature*, **165**, 240.
Du Mond, J. W. M. & Cohen, E. R. 1948 *Rev. Mod. Phys.* **20**, 82.
Dymott, E. R. & Jennings, W. O. 1948 *Machinery* (21 Oct.), p. 590.
Essen, L. 1946 *J. Instn Elect. Engrs*, **93**, pt. 3A, no. 9, p. 1413.
Essen, L. 1947 *Nature*, **159**, 611.
Essen, L. 1950 *Nature*, **165**, 582.
Essen, L. & Gordon-Smith, A. C. 1948 *Proc. Roy. Soc. A*, **194**, 348.
Houstoun, R. A. 1949 *Nature*, **164**, 1004.
Husten, B. F. & Lyons, H. 1948 *Trans. Amer. Instn Elect. Engrs*, **47**, pt. 1, p. 321.
Whinnery, J. R. 1942 *Electronics* (Feb.), p. 44.

A critical examination of the gravitational theory of the origin of cosmic rays

BY P. J. D. GETHING

Department of Mathematics, Imperial College of Science, London

(Communicated by E. A. Milne, F.R.S.—Received 18 July 1950)

The observational data which any theory of the origin of primary cosmic radiation should explain and the comparative merits of the various theories so far propounded are examined. In particular, Milne's suggestion concerning a possible gravitational acceleration in intergalactic space is considered in detail, and his somewhat involved analysis is replaced by a simpler method of derivation by means of an alternative time-scale. The energy spectrum and possible total intensity of gravitationally accelerated material are examined; it is shown that an expression can be obtained for the former which is consistent with cosmic-ray data, but because the world-model considered is not sufficiently sharply defined, no order of magnitude can be obtained for the latter. It thus appears that a gravitational theory of cosmic radiation can account for most of the main features of the observational data, and as regards the remaining features, although the present theory is incomplete, it presents no actual inconsistencies with observation.

1. INTRODUCTION

Any theory of origin of cosmic rays should be capable of accounting for:

- (i) the high energies of the individual particles,
- (ii) the observed total intensity,
- (iii) the energy spectrum,
- (iv) the isotropy,
- (v) the constitution of the radiation as observed at the earth.

The energies of the individual particles range from a lower limit at about 2×10^9 up to 10^{14} eV or even higher. The total energy flux across 1 sq.cm. in 1 sec., after the screening effect of the earth's atmosphere has been allowed for, is found to be about 3×10^{-3} erg. This implies an energy density in the neighbourhood of the earth of the order of 10^{-13} erg cm.⁻³.

If $n(E)dE$ denotes the number of particles in the primary radiation with energies in the range E to $E+dE$, then provided that E is greater than 2×10^9 eV a good approximate formula for the energy spectrum is given by

$$n(E)dE = \text{const. } dE/E^\gamma, \quad (1)$$

where γ is a constant of value about 2.8. The absence of cosmic radiation with energies below the value quoted, the so-called 'cut-off' value, has been attributed by Janossy (1937) to the deflecting action of the solar magnetic field. Recent observations, however, have suggested that this theory is untenable, and an alternative explanation may be needed. For example, the results obtained by Thiessen (1949) for the sun's magnetic moment are not consistent with the value required by Janossy's theory (see Gething 1950). Dolbear & Elliot (1950a) have recently reported that the diurnal variations in intensity are of the order of 0.02 %, and do

not increase significantly in amplitude with increasing altitude (1950*b*), whereas Janossy's theory predicts rather larger variations increasing to a few percent at high altitudes. Moreover, Pomerantz (1950) has obtained evidence that the cut-off is not sharp and that some radiation is reaching the earth below the critical value. The exponent γ of the power law (1) appears to decrease near the low-energy end of the spectrum (Pomerantz & Vallarta 1949).

The observed constant intensity at any given point on the earth's surface (apart from very small fluctuations of uncertain significance) shows that the intensity is independent of the direction of the observer's zenith. Hence, it would appear that the radiation arrives equally from all directions. This apparent isotropy of the primary radiation is one of the outstanding features of the observational data and presents a crucial test for any theory.

The primary radiation is believed to consist largely of protons together with a small proportion of α -particles and heavy nuclei, but the possibility of the presence of uncharged particles cannot be excluded. Bradt & Peters (1950) estimate that the proportions of protons, α -particles and heavy nuclei reaching the earth are in approximately the same ratio as the respective cosmical abundances of hydrogen, helium and the metals of low atomic weight.

2. THEORIES OF ORIGIN

The problem of finding a suitable mechanism capable of giving individual particles the high energies which are observed is by no means easy, and a wide variety of possible sources of cosmic radiation has been suggested. Only the more important theories will be examined here.

(i) Fermi (1949) has suggested that the cosmic-ray particles acquire their high energies from 'collisions' with magnetic fields associated with drifting interstellar clouds. This theory accounts satisfactorily for the high individual energies of the particles; in addition it is possible to derive an energy spectrum consistent with observation, by making the reasonable assumption that the mean distance which a proton travels between two collisions of this type is of the order of one light-year. Apparently no attempt has yet been made on the basis of this theory to account for the observed isotropy and total intensity. However, the main difficulty of the theory is that the initial energy of a proton must be at least 2×10^8 eV before it can be accelerated by this mechanism. For lower initial energies the possible energy increments will be exceeded by the losses due to ionization. Consequently, some additional mechanism is needed to bring the particles up to this 'injection energy'. The initial energies of heavy nuclei would have to be even higher, whereas observations by Hoang (1950) show that a considerable proportion of the heavy particles reaching the earth have energies *below* the appropriate value calculated on the basis of Fermi's theory.

(ii) Davis (1947) has suggested that rotating magnetized stars might be a possible source of cosmic rays. Since rotation of the star through its magnetic field gives rise to an electric field, the stellar surface acquires a potential with respect to the surrounding space. Charged particles will leave the surface and a space charge will

be built up, tending to neutralize the field. Babcock (1948) has pointed out that, for stars in which the sense of magnetization changes periodically, the suppressing effect of the space charge would be absent at certain phases. According to the limited investigations made on the basis of this theory it appears that the mechanism considered may be capable of generating energies up to 10^{14} eV, and this range is not incompatible with the at present somewhat meagre observational data.

The following objections can be brought to bear against this theory:

(a) Davis himself pointed out a serious difficulty. The energies of the particles must be acquired at the expense of the rotational energy of the star; estimates of the total rotational energy of all the stars in the galaxy suggest that it is insufficient to maintain the cosmic radiation at its present intensity for more than a few hundred years. However, this argument loses its force if cosmic rays are assumed to be completely imprisoned in the galaxy by a magnetic field, as was originally suggested by Alfvén (1937).

(b) It is also difficult to account for the observed isotropy, because the suggested sources would presumably be strongly concentrated towards the galactic plane. It is, however, possible that the magnetic fields associated with these supposed stellar sources may be capable of deflecting the charged particles at large distances so as to distribute them more isotropically.

Hoyle (1946) also has put forward a theory relating cosmic rays to stellar rotation. He maintains that gravitationally collapsing stars become rotationally unstable and eject particles with extremely high speeds. This theory depends on the hypothesis that stars of this type exist.

(iii) Ryle (1949) also has suggested a possible galactic origin. He associates the sources of cosmic radiation with the recently discovered discrete sources of radio noise. The mean kinetic energy of particles in the latter is believed to be of the order of 10^{10} eV at least, which would be sufficient to account for the observed energies of most cosmic-ray particles.

With regard to this theory, the deflecting action of magnetic fields is again needed to explain the observed isotropy. It is necessary to suppose that the fields associated with the sources are causing appreciable deflexions at distances at least a million times greater than the linear dimensions of the sources if these are comparable in size with ordinary stars; even so, the distribution of the highest energy cosmic-ray particles might not be strictly isotropic.

(iv) Richtmeyer & Teller (1949) have pointed out that many of the difficulties mentioned above would be avoided if cosmic rays were a purely local phenomenon with a solar origin. Alfvén (1949, 1950) has developed this idea and suggested a possible mechanism by which charged particles moving in orbits round the sun could gain energy from a postulated changing dipole field. His calculations indicate that the energies of the particles, the total intensity and the energy spectrum might all be explained on the basis of this theory. However, an isotropic distribution at the earth can only be obtained by postulating *ad hoc* a magnetic field of particular 'shape' and intensity. Alfvén suggests that such a field might arise from the motion of the solar system through magnetized interstellar matter. There is, however, no observational evidence so far for its existence.

The theories considered above have all concerned possible sources within the galaxy, postulating either a stellar, interstellar or solar origin. An extra-galactic origin has been considered in other theories, notably those of Lemaître and of Milne.

(v) Lemaître (1949) regards cosmic rays as fundamental constituents of the universe. He believes that the universe has evolved by a process of repeated division from one gigantic particle, 'l'atome primitif'. In the process of division, cosmic-ray particles are produced. Thus the difficulties of explaining the origin of cosmic radiation are relegated to the fundamental problem of cosmogony. The theory cannot be subjected to any quantitative test, but it is interesting to note that Lemaître (1931) predicted a component of heavy nuclei in the primary radiation long before the discovery of such particles.

(vi) Milne (1935) has suggested that cosmic radiation is primarily due to world gravitation. He has suggested that particles in inter-galactic space could be accelerated by the combined gravitational field of all the nebulae, and thus eventually attain velocities arbitrarily close to the velocity of light in finite time and distance. This theory accounts easily and naturally for the high energies of the cosmic-ray particles and their isotropic distribution. In addition, George (1948) has derived an energy spectrum of the right form. Unfortunately:

(a) The idealized model of the universe adopted by Milne involves infinite density singularities at the fundamental particles or nuclei identified with the extra-galactic nebulae. This is an imperfection in the model; however, there is no reason to suppose that the same phenomenon could not arise if the singularities were removed, or even if an entirely different model of the universe were adopted.

(b) No numerical value can be assigned to the intensity to be expected. This point is considered in greater detail in §9.

It will be seen that each of the above theories either fails to explain all the observational data or is unsatisfactory in some other respects, and in fact none has yet been generally accepted. However, the gravitational theory of Milne appears to be in many ways the most promising, and it is the purpose of this paper to give a simplified treatment of his analysis. In view of the somewhat involved method originally followed by Milne it is important to test the validity of his conclusions and establish the criteria on which they depend. In this way a clearer understanding of the physical assumptions and results involved may be obtained.

3. PHYSICAL BACKGROUND OF THE GRAVITATIONAL THEORY

In his books *Relativity, gravitation and world structure* (1935) (referred to throughout this paper as WS) and *Kinematic relativity* (1948) (KR), Milne has examined in great detail the properties of a particular class of world-models. The most primitive and fundamental of these he calls the 'substratum'. Essentially this is a homogeneous, isotropic fluid which can be regarded as in uniform radial expansion from each constituent particle, according to a law of 'Hubble's type' $V = r/t$, where V denotes velocity at epoch t of a particle at distance r . With each particle can be associated a 'fundamental observer', and the particles can be distinguished from others introduced later by the term 'fundamental particle'. These fundamental observers

are provided with congruent clocks and by signalling experiments (KR, chap. II) they can determine, in principle, the relative motions, etc., of all other particles. Under certain conditions governing the geometry of the system, supplementing the homogeneity and isotropy conditions already mentioned, each observer will map the density distribution of the whole system according to the law

$$n dx dy dz = B t dx dy dz / c^3 (t^2 - \mathbf{r}^2/c^2)^2.$$

In addition to the fundamental particles, we can consider free test-particles whose motion is controlled by the system of fundamental particles. The equation of motion of a free-particle was found by Milne to be of the form

$$\frac{d\mathbf{V}}{dt} = (\mathbf{r} - \mathbf{V}t) \frac{Y}{X} G(\xi), \quad (2)$$

where $\mathbf{r}$ and $\mathbf{V}$ denote position and velocity at epoch t , $X \equiv t^2 - \mathbf{r}^2/c^2$, $Y \equiv 1 - \mathbf{V}^2/c^2$ and $\xi \equiv Z^2/X Y$, where $Z \equiv t - \mathbf{r} \cdot \mathbf{V}/c^2$. $G(\xi)$ is a function of ξ which does not appear at first to be uniquely specified. However, equation (2) from its manner of derivation is also the equation of motion of a free-particle in any dynamical system based on the substratum which appears to have the same properties according to each fundamental observer. Milne has shown that there are other systems of this type, namely systems formed by superposing on the substratum a statistical distribution of subsidiary particles governed by the law

$$N(\mathbf{r}, \mathbf{V}, t) dx dy dz du dv dw = \frac{\psi(\xi) dx dy dz du dv dw}{c^6 X^\dagger Y^\dagger}, \quad (3)$$

where $\psi(\xi)$ can be any function of ξ .

The function $G(\xi)$ occurring in the formula (2) for the acceleration can be related to the function $\psi(\xi)$ defining the statistical distribution by means of a generalized Boltzmann equation, expressing the condition that particles are assumed to be neither created nor destroyed. This equation can be written in the form

$$\frac{\partial N}{\partial t} + \mathbf{V} \cdot \frac{\partial N}{\partial \mathbf{r}} + \frac{d\mathbf{V}}{dt} \cdot \frac{\partial N}{\partial \mathbf{V}} + N \left(\frac{\partial}{\partial \mathbf{V}} \cdot \frac{d\mathbf{V}}{dt} \right) = 0, \quad (4)$$

and yields, on substituting (3) and integrating, the general relation

$$\{G(\xi) + 1\} \psi(\xi) = -C/(\xi - 1)^\dagger. \quad (5)$$

It is clear that in general, given $\psi(\xi)$, then, to within an arbitrary constant C , the function $G(\xi)$ is uniquely defined. Conversely, corresponding to a given function $G(\xi)$, then, to within an arbitrary constant C , the function $\psi(\xi)$ is uniquely determined. This one-one relationship persists except when $\psi(\xi) = 0$ or $G(\xi) = -1$. Corresponding to all other $\psi(\xi)$, we have $G(\xi) \neq -1$ unless $C = 0$, and to all other $G(\xi)$ we have $\psi(\xi) \neq 0$ unless $C = 0$. Milne (KR, p. 133) has argued that when $\psi(\xi) = 0$ the function $G(\xi)$ necessarily takes the value -1 . Consequently, the acceleration of a free-particle in the substratum is governed by the equation

$$\frac{d\mathbf{V}}{dt} = -(\mathbf{r} - \mathbf{V}t) \frac{Y}{X}, \quad (6)$$

whereas in the case of a substratum with a superposed statistical component determined by (3) the acceleration of a free particle is given by (2), where $G(\xi)$ is given by (5).

It is clear that both the pure substratum and a substratum with a superposed statistical component of the type considered can at best represent idealized world-models suitable for investigating features of the universe which are not primarily determined by local irregularities. If the pure substratum is regarded as a first large-scale approximation to the astronomical universe, providing a uniform expanding material 'background', we can regard a system with a suitably chosen statistical component as a somewhat more sophisticated model which may have properties that can be associated with certain physical phenomena for the elucidation of which the pure substratum is too crude a model to use. We can visualize a hierarchy of abstract systems, beginning with the substratum and arranged in order of increasing complexity with the actual astronomical universe, with all its local irregularities, etc., as an ultimate limiting form. A substratum with a statistical component would then represent the second level in this hierarchy. It reproduces in a general way the feature of clustering around nuclei, idealized as particles and themselves uniformly distributed, which characterizes the extra-galactic nebulae at a certain level of simplification.

Assuming that phenomena such as the nebular red-shifts can be associated, in the first instance at least, with the crudest of our world-models, it is natural to look for possible phenomena which can be associated in their most basic form with the second type of model in our hierarchy. We have already remarked that the primary radiation incident at the top of the atmosphere is spatially isotropic. This fact suggests that if the primary* origin of the radiation is to be sought in properties of the world as a whole, then the appropriate model to consider is a uniform isotropic one. It has been shown by Whitrow (1940) that there is no intrinsic acceleration of a free-particle in the presence of the substratum alone. Milne, however, claims that a free-particle projected in a substratum with a superposed non-vanishing statistical component will acquire an intrinsic acceleration, so that if it is not deflected in any way it will eventually reach a speed arbitrarily close to that of light.

The fundamental particles (idealized nebular nuclei) do not experience this acceleration themselves, but move with uniform relative velocities. Each observer attached to such a particle considers himself to be at a centre of symmetry of the universe. An observer attached to a free-particle projected with a non-zero relative velocity from a fundamental particle would regard the centre of symmetry of the universe to be at a finite non-zero distance away in the direction of his motion relative to the fundamental particle from which he is projected. The greater the velocity of projection of the free-particle, the more distant this centre will appear to be, tending to the confines of the system for velocities close to c . The free-particle falls towards this apparent centre, but as its velocity increases so its distance from the centre also increases and hence the velocity must increase still further. In general, except when $G(\xi) \equiv -1$, Milne claims that a velocity arbitrarily close to c can be generated in a finite time. In the limiting case of a pure substratum, $G(\xi) = -1$, the relative velocity with respect to the original observer will tend to c only after an infinite time, although the local velocity (i.e. relative to the fundamental particle which the free-particle is passing) will be invariant. Consequently, if a world-gravitational origin is

* There may be additional secondary origins accounting for fluctuations and anomalies.

sought for the high energies of cosmic-ray particles it must be associated with a substratum on which a statistical component has been superposed, or with some more complicated system. If Milne's results are correct, then it is suggested that the primary origin of cosmic-ray energies is a consequence of the tendency of matter to cluster about nebular nuclei.

4. AN ALTERNATIVE MATHEMATICAL APPROACH TO THE GRAVITATIONAL THEORY: TRANSFORMATION OF TIME-SCALE

It is well known (see KR, p. 57) that a regraduation of the observer's clock from the t -scale to a new scale T given by

$$T = t_0 \log (t/t_0) + t_0, \quad (7)$$

where T is the label of the instant previously labelled t , reduces the uniformly expanding substratum to a stationary distribution of constant density. The model of the universe, which previously appeared to the observer to be bounded by a sphere of radius $r = ct$, now becomes of infinite extent. An event (r, t) transforms to an event (R, T) given by

$$t = t_0 \exp \{(T - t_0)/t_0\} \cosh (R/ct_0), \quad (8)$$

$$r = ct_0 \exp \{(T - t_0)/t_0\} \sinh (R/ct_0). \quad (9)$$

t_0 is a constant defining the epoch at which the times T and t agree.

For a given free-particle, a subclass of fundamental observers can always be found such that the velocity vector and the radius vector of the free-particle relative to any member of this subclass coincide in direction, i.e. such that the motion is in the line of sight. Then the acceleration of the particle relative to such an observer must also be directed along this line, because of the symmetry of the model.

The scalar relation between a velocity V measured in (r, t) co-ordinates and the corresponding velocity U measured in (R, T) co-ordinates is (Whitrow 1939)

$$\tanh^{-1}(V/c) = \tanh^{-1}(U/c) + R/ct_0. \quad (10)$$

The following relations will also be needed. From (8) and (9),

$$t^2 - r^2/c^2 = t_0^2 \exp \{2(T - t_0)/t_0\},$$

$$\text{i.e.} \quad X^{\frac{1}{2}} = t_0 \exp \{(T - t_0)/t_0\}. \quad (11)$$

From equation (2) and the definitions of X , Y , Z , ξ , we have

$$\frac{dX}{dt} = 2Z, \quad \frac{dY}{dt} = -2Y(tY - Z)G/X, \quad \frac{dZ}{dt} = Y[1 + (X - tZ)G/X]. \quad (12)$$

$$\text{Hence} \quad \frac{d\xi}{dt} = -\frac{2Z}{X}(\xi - 1)(1 + G), \quad (13)$$

a result previously obtained by Milne (see WS, p. 143). It is, however, a relation independent of the form of the function G . Milne uses an argument based on the

theory of dimensions (Note 3, WS, p. 351) to show that G is a function of ξ only, and cannot depend on X . Differentiation of (11) gives

$$\frac{dX}{dT} = \frac{2X}{t_0}, \quad (14)$$

which, with (12), becomes

$$\frac{dT}{dt} = \frac{t_0 Z}{X}. \quad (15)$$

Whitrow (1940) has proved that the velocity of a particle passing an observer, as measured by the observer, is independent of the time-scale he employs. This result enables us to obtain the equation

$$\xi = 1/(1 - U^2/c^2) \quad (16)$$

(see KR, p. 79) by elementary reasoning as follows. At the observer's origin, $\mathbf{r} = 0$, the expression for ξ reduces to $\xi = 1/(1 - V^2/c^2)$ in (r, t) co-ordinates. In virtue of the result quoted above, this must transform at the origin to $\xi = 1/(1 - U^2/c^2)$. Since on the T -scale the fundamental particles are at rest relative to each other, the relation (16) must still be valid when the free-particle is at a finite distance from the observer. Hence the relation holds generally.

5. FREE-PARTICLE TRAJECTORIES

It is now possible to derive the differential equation of motion in T -time for a free-particle corresponding to equation (2). This is most simply obtained by use of the identity

$$\frac{dU}{dT} = \left(\frac{dU}{d\xi} \right) \left(\frac{d\xi}{dt} \right) \left(\frac{dt}{dT} \right). \quad (17)$$

Differentiation of (16) gives immediately

$$\frac{dU}{d\xi} = \frac{c}{2(\xi - 1)^{\frac{1}{2}} \xi^{\frac{3}{2}}}. \quad (18)$$

Substitution of (13), (15) and (18) in (17) gives

$$\frac{dU}{dT} = \frac{c}{2(\xi - 1)^{\frac{1}{2}} \xi^{\frac{3}{2}}} \left\{ -\frac{2Z}{X} (\xi - 1) (1 + G) \right\} \frac{X}{t_0 Z},$$

$$\text{i.e.} \quad g(U, T) = -(c/t_0) (\xi - 1)^{\frac{1}{2}} \xi^{-\frac{3}{2}} (1 + G), \quad (19)$$

where $g(U, T)$ is the new acceleration function.

In the substratum, G reduces identically to -1 and equation (19) shows immediately that the acceleration of a free-particle, as measured on the T -scale, must then be zero. Thus the free-particle, if originally set in motion by some external agency, will continue to move with uniform velocity. The velocity of light cannot be reached in finite time or distance in (R, T) co-ordinates, confirming the result quoted in §3.

Another consequence of equation (19) is that if G is a function of ξ only, dU/dT is also a function of ξ only. Equation (16) shows that this may be written as a function $g(U)$ of U only. The functional identity for g in terms of G is

$$g(U) \equiv -\frac{U}{t_0} \left(1 - \frac{U^2}{c^2}\right) \left[1 + G\left(\frac{1}{1 - U^2/c^2}\right)\right]. \quad (20)$$

If the function G includes X as a variable, g will depend on T , but not on R :

$$g(U, T) \equiv -\frac{U}{t_0} \left(1 - \frac{U^2}{c^2}\right) \left[1 + G\left(t_0^2 \exp\{2(T - t_0)/t_0\}, \frac{1}{1 - U^2/c^2}\right)\right]. \quad (21)$$

The former case corresponds to the problem treated by Milne (WS, chap. VIII). By transforming the time-scale and considering the acceleration relative to an observer in the line of motion, the equation of a free-particle trajectory has been reduced to the comparatively simple scalar equation

$$\frac{dU}{dT} = g(U). \quad (22)$$

6. SOLUTION OF THE TRAJECTORY EQUATIONS

The solution of equation (22), together with the equation defining U ,

$$\frac{dR}{dT} = U, \quad (23)$$

will now be obtained and compared with the solutions deduced by Milne. Again, it must be emphasized that the problem has been rendered effectively one-dimensional by choosing the observer to be in the line of motion, without, however, any real loss of generality.

Suppose the initial conditions are denoted by (R_1, T_1, U_1) . The corresponding values in terms of the t -scale will be denoted by (r_1, t_1, V_1) . Milne takes r_1 and t_1 to be zero, i.e. he assumes that the free-particle starts its career at the creation of the universe. Creation is relegated to the infinite past on the T -scale, as may be seen by substituting $t = 0$ in equation (7). In discussing whether particles can accelerate to the velocity of light in finite time in terms of the T -scale it would be pointless to assume that they became free-particles infinitely long ago. It will therefore be necessary to assume that a particle may be ejected from a nucleus by some subsidiary mechanism at an instant such that $t_1 \neq 0$ and hence the epoch measured on the T -scale, T_1 , is finite. The relations between R_1 and r_1 , T_1 and t_1 , U_1 and V_1 are given by equations (9), (8) and (10) respectively.

The initial acceleration was taken to be in the direction $-\mathbf{i}$ by Milne, where $\mathbf{i}$ is a unit vector whose direction is specified by two constants of integration. In the one-dimensional case, $\mathbf{i}$ must be replaced by -1 in the solutions. The parameter ζ , defined by the relations

$$\sinh \zeta = +(\xi - 1)^{\frac{1}{2}}, \quad \cosh \zeta = \xi^{\frac{1}{2}},$$

is now seen to have a new significance, for equation (16) gives

$$\tanh \zeta = U/c. \quad (24)$$

The definition of the initial constant ϵ simplifies to

$$\sinh \epsilon = -\frac{V_1/c}{(1 - V_1^2/c^2)^{\frac{1}{2}}}, \quad \tanh \epsilon = -V_1/c. \quad (25)$$

Substitution of equation (10) into the definition of Y , i.e.

$$Y = (1 - V^2/c^2),$$

gives, with (24),

$$Y = 1/\cosh^2(\zeta + R/ct_0). \quad (26)$$

Milne has given (WS, p. 153) vector solutions to the initial trajectory equations

$$\frac{d\mathbf{r}}{dt} = \mathbf{V}, \quad \frac{d\mathbf{V}}{dt} = (\mathbf{r} - \mathbf{V}t) \frac{Y}{X} G(\xi). \quad (27)$$

These solutions will now be transformed and simplified for the particular case considered here. It has already been shown that the trajectory equations (27) transform to

$$\frac{dR}{dT} = U, \quad \frac{dU}{dT} = g(U). \quad (28)$$

It will be confirmed that the transforms of the solutions obtained by Milne for equations (27) do in fact satisfy equations (28)

Thus Milne's equation (51)

$$\mathbf{V} = \mathbf{V}_1 - i c (1 - V_1^2/c^2)^{\frac{1}{2}} \sinh(\eta + \zeta) / \cosh(\epsilon - \eta - \zeta)$$

becomes
$$\tanh(\zeta + R/ct_0) = -\tanh \epsilon + \frac{\sinh(\eta + \zeta)}{\cosh \epsilon \cosh(\epsilon - \eta - \zeta)},$$

or
$$\frac{\sinh(\epsilon + \zeta + R/ct_0)}{\cosh(\zeta + R/ct_0)} = \frac{\sinh(\eta + \zeta)}{\cosh(\eta + \zeta - \epsilon)}. \quad (29)$$

Milne's equation (52)

$$\mathbf{r} = \mathbf{V}_1 t - i c X^{\frac{1}{2}} (\sinh \eta) / \cosh \epsilon$$

becomes
$$\sinh(R/ct_0) = -\tanh \epsilon \cosh(R/ct_0) + (\sinh \eta) / \cosh \epsilon,$$

or
$$\sinh(\epsilon + R/ct_0) = \sinh \eta. \quad (30)$$

Milne's equation (53)

$$t = \frac{X^{\frac{1}{2}} \cosh(\epsilon - \eta)}{(1 - V_1^2/c^2)^{\frac{1}{2}} \cosh \epsilon}$$

becomes
$$\cosh(R/ct_0) = \cosh(\epsilon - \eta). \quad (31)$$

Milne's equation (55)

$$Y^{\frac{1}{2}} = (1 - V_1^2/c^2)^{\frac{1}{2}} \frac{\cosh \epsilon}{\cosh(\epsilon - \eta - \zeta)}$$

becomes
$$\cosh(\zeta + R/ct_0) = \cosh(\epsilon - \eta - \zeta). \quad (32)$$

Milne's definition of the X -integral, equation (56),

$$X = X_1 \exp \left[- \int_{\xi_1}^{\xi} \frac{d\xi}{(\xi - 1) \{1 + G(\xi)\}} \right]$$

becomes
$$\exp \{2(T - t_0)/t_0\} = \exp \{2(T_1 - t_0)/t_0\} \exp \left\{ \frac{2}{t_0} \int_{U_1}^U \frac{dU}{g(U)} \right\},$$

i.e.
$$T - T_1 = \int_{U_1}^U \frac{dU}{g(U)}. \quad (33)$$

The definition of η , Milne's equation (57),

$$\eta = -\frac{1}{2} \int_1^\xi \frac{d\xi}{\xi^{\frac{1}{2}}(\xi-1)^{\frac{1}{2}}\{1+G(\xi)\}}$$

becomes

$$\eta = \frac{1}{ct_0} \int_0^U \frac{UdU}{g(U)}. \quad (34)$$

Equations (29), (30), (31) and (32) above all lead to the same solution for η ,

$$\begin{aligned} \eta &= \epsilon + R/ct_0 \\ &= (R - R_1)/ct_0 - \tanh^{-1}(U_1/c), \end{aligned}$$

which, with (28), gives

$$\eta = \frac{1}{ct_0} \int_{U_1}^U \frac{UdU}{g(U)} - \tanh^{-1}(U_1/c). \quad (35)$$

The reason for the extra term in (35), as compared with (34), is the difference in the initial conditions assumed in deriving these two equations. If the particle starts its career with a non-zero velocity U_1 the integral occurring in (34) is not defined between the limits 0 and U_1 . η must be defined by equation (35), and equation (34) regarded as the limiting case as $U_1 \rightarrow 0$. With this stipulation, it will be seen that all the equations are consistent and more particularly that (33) is a solution of (28). The results obtained by Milne are thus verified.

7. IDENTIFICATION OF FREE-PARTICLES WITH COSMIC RAYS

The results of the previous section show that a free-particle will reach the velocity of light at a distance and time given by

$$R = R_1 + \int_{U_1}^c \frac{UdU}{g(U)}, \quad (36)$$

$$T = T_1 + \int_{U_1}^c \frac{dU}{g(U)}, \quad (37)$$

respectively. The somewhat involved proof given by Milne that each of these quantities is finite (i.e. that the integrals in equations (36) and (37) are convergent) will now be replaced by a simpler proof based on the method developed above, thereby verifying Milne's results.

First the relation between $\psi(\xi)$ and $g(U)$ is needed. Substitution of (20) into (5) gives, on simplification,

$$\psi \left(\frac{1}{1 - U^2/c^2} \right) = \frac{Cc^3(1 - U^2/c^2)^{\frac{1}{2}}}{t_0 g(U) U^2}. \quad (38)$$

It is now possible to transform Milne's expressions for the density at the observer's origin of particles moving with speeds between specified limits (WS, p. 188, equation (5)) and for the total number of particles striking unit area at the origin in unit time (WS, p. 190, equation (6)). It is only necessary to note that, at the origin, velocities and velocity ranges have the same values irrespective of the particular time-scale employed. Also, the space may be treated as Euclidean as a local approximation, so

that the volume element remains $dx dy dz$. Then the number of particles at the origin moving with speeds between V_1 and c ,

$$\frac{4\pi dx dy dz}{c^6 t^3} \int_{V_1}^c \psi \left(\frac{1}{1 - V^2/c^2} \right) \frac{V^2 dV}{(1 - V^2/c^2)^{\frac{1}{2}}} \quad (39)$$

becomes, in virtue of (8) and (38),

$$\frac{4\pi C dx dy dz}{c^3 t_0^4 \exp \{3(T - t_0)/t_0\}} \int_{U_1}^c \frac{dU}{g(U)}. \quad (40)$$

The number of particles with speeds between V_1 and c impinging on an area dS at the origin in an interval dt is given by

$$\frac{\pi dt dS}{c^6 t^3} \int_{V_1}^c \psi \left(\frac{1}{1 - V^2/c^2} \right) \frac{V^3 dV}{(1 - V^2/c^2)^{\frac{1}{2}}},$$

which becomes, for the corresponding interval dT ,

$$\frac{\pi C dT dS}{c^3 t_0^4 \exp \{2(T - t_0)/t_0\}} \int_{U_1}^c \frac{U dU}{g(U)}. \quad (41)$$

The significance of these expressions (40) and (41) is due to the fact that they contain the same integrals as those occurring in (37) and (36) respectively. The model so far adopted involves both a function and a constant which have not been specified, so that some degree of freedom is left in the choice of $\psi(\xi)$ and hence of the function $g(U)$; however, it is apparent that if direct comparison is possible between the present model and astronomical observation, then the number of particles in the fundamental observer's neighbourhood with speeds between a *non-zero* value U_1 and the speed of light c , must be finite, as must the number of impinging particles with speeds in the same range. Hence the integrals concerned must converge, and from equations (36) and (37) it follows that free-particles leaving a nucleus will be seen to be accelerated to a speed arbitrarily close to c in a finite distance and finite time. *More generally, if an observer finds that the density of particles with speeds between U_1 and U_2 ($U_2 > U_1$) is finite, he will also observe particles accelerating from U_1 to U_2 in finite time.* Milne has already noted that there is a singularity (infinite density) when $\xi = 1$, i.e. when $U_1 = 0$, which must be associated with the idealization of each nebular nucleus in the form of a 'particle' with vanishing linear dimensions. We now see that this feature automatically ensures that a free-particle cannot be accelerated from rest at a nucleus, in accordance with the condition that the nucleus (fundamental particle) is itself a free-particle.

Milne has suggested that free-particles, accelerated to speeds very close to c by the mechanism considered above, may be identified with primary cosmic radiation.

8. MOTION OF A FREE-PARTICLE AFTER REACHING THE VELOCITY OF LIGHT

For the sake of completeness it is now necessary to consider the motion of a free-particle after it has attained the velocity of light. The velocity must be assumed to begin to decrease immediately, in order that the probability of an observer meeting a material particle whose velocity is exactly c is zero. The number of material particles in the neighbourhood of the (fundamental) observer at a given time with

speeds U in any given range $0 < U_1 < U < c$ is finite, but approaches zero as U_1 approaches c .

Milne has argued (WS, p. 236) that the change from acceleration to deceleration postulated above is associated with the fact that the function $g(U)$ is double-valued, since equation (38) contains a square-root; the positive value, corresponding to an acceleration, being taken along the first branch of the trajectory until the velocity of light is reached, and the negative value along the remainder of the path. This procedure can be justified as follows. We have already noted that as the speed of a free-particle approaches c , so the apparent centre of symmetry of the whole system to an observer rigidly attached to the particle approaches the confines of the system. On the T -scale, the world-model here considered is of infinite extent. Consequently the apparent centre 'approaches infinity' in the direction of motion of the particle. The time interval between the instant of ejection of the free-particle from the nucleus and the critical instant at which the free-particle reaches the speed of light is finite. Just before this critical instant the apparent centre is very distant in the forward direction of the particle's motion. Mathematical continuity suggests that immediately afterwards the apparent centre should be very distant in the rearward direction of the particle's motion. This agrees with the condition that immediately after the critical instant the particle tends to slow down, the 'pull of the universe' now acting as a force of deceleration. Although the particle appears now to be moving away from the apparent centre all the time its distance from this centre is decreasing. As Milne has argued, the free-particle continues to decelerate until it is brought to rest, when an observer attached to it again judges himself to be at the centre of the universe.

The complete career of a free-particle can now be traced. If it is ejected from a fundamental particle at a distance R_1 from the observer, with a velocity U_1 in the line of sight, at a time T_1 , it will accelerate until it reaches the velocity of light at a finite distance given by (36) and a finite epoch given by (37). From this point onward the particle will decelerate, the velocity decreasing to its initial value U_1 at a finite distance

$$R = R_1 + 2 \int_{U_1}^c \frac{U dU}{g(U)}$$

and a finite epoch

$$T = T_1 + 2 \int_{U_1}^c \frac{dU}{g(U)}.$$

It will continue to decelerate towards a zero velocity, but it cannot come to rest in a finite time. The limiting position which it approaches is, however, at a finite distance

$$R = R_1 + 2 \int_{U_1}^c \frac{U dU}{g(U)} + \int_0^{U_1} \frac{U dU}{g(U)}.$$

Thus the free-particle will tend to be 'captured' by another fundamental particle at this distance from the observer. Particles ejected from a given fundamental particle with equal initial velocities in the same direction, but at differing times, will approach a common limiting position and form themselves into a sub-system. Milne has suggested the possibility (WS, chap. XIII) that sub-systems so generated may be analogous to the interstellar clouds of particles which are present in our own galaxy and probably in others.

9. ENERGY SPECTRUM AND INTENSITY

The condition that $g(U)$ must be zero when $U = c$ imposes a restriction on the form of $\psi(\xi)$. Equation (38), when rewritten in terms of ξ , shows that $\xi^{\frac{1}{2}}\psi(\xi) \rightarrow \infty$ as $\xi \rightarrow \infty$. Milne has also shown that, since the integral in (39) must converge as $V_1 \rightarrow c$, $\psi(\xi)$ for large ξ must be less than $\text{const. } \xi^{-\frac{1}{2}}$. These two conditions are summarized in his equation (8) (WS, p. 191),

$$\frac{N}{\xi^{\frac{1}{2}}} < \psi(\xi) < \frac{\text{const.}}{\xi^{\frac{1}{2}}} \quad (\xi \sim \infty), \quad (42)$$

where N is an arbitrary large constant.

There is another physical condition which it seems reasonable to impose so as to limit the form of $\psi(\xi)$ still further, namely, that the total energy carried by all the free-particles in a given finite volume element is finite. The energy of a free-particle is proportional to $\xi^{\frac{1}{2}}$, and its numerical value is independent of the choice of observer, so that we can consider a volume element $dx dy dz$ at the observer's origin without loss of generality. At the origin

$$\xi^{\frac{1}{2}} = (1 - V^2/c^2)^{-\frac{1}{2}},$$

and, with the help of expression (39), the condition is seen to be that the integral

$$\int_{V_1}^c \psi \left(\frac{1}{1 - V^2/c^2} \right) \frac{V^2 dV}{(1 - V^2/c^2)^3}$$

should converge as $V_1 \rightarrow c$. This implies a new upper limit for $\psi(\xi)$,

$$\psi(\xi) < \text{const.}/\xi^2. \quad (43)$$

By assuming that $\psi(\xi)$ may be written approximately as $A/\xi^{\frac{1}{2}}$ for large ξ , where A is a constant in the range considered, George (1948) has succeeded in obtaining a theoretical energy spectrum for the particles reaching the observer, i.e.

$$n(E) dE = \pi A m_0^2 c^2 (1 - m_0^2 c^4/E^2) dE/t^3 E^3,$$

where m_0 is the mass of a free-particle. For $E \gg m_0 c^2$ this becomes approximately

$$n(E) dE = \text{const. } dE/E^3,$$

in good agreement with the experimental curve (1) for cosmic rays reaching the earth.

However, examination of (42) and (43) suggests the use of the more general form for $\psi(\xi)$,

$$\psi(\xi) = A'/\xi^{(1-\alpha)},$$

where α must lie in the range $0 < \alpha < \frac{1}{2}$, for large ξ ; α and A' are constants in the range considered. With this modification the energy spectrum becomes

$$n(E) dE = \pi A' m^{(2-2\alpha)} c^{(2-4\alpha)} (1 - m_0^2 c^4/E^2) dE/t^3 E^{(3-2\alpha)}. \quad (44)$$

Then for $E \gg m_0 c^2$ the curve approximates to

$$n(E) dE = \text{const. } dE/E^{(3-2\alpha)}, \quad 0 < \alpha < \frac{1}{2}$$

The experimental cosmic-ray curve is subject to considerable uncertainties, but it appears that if α were about 0.1 or 0.2 the theoretical curve would be in even better agreement with observation than it is in the particular case chosen by George, $\alpha = 0$. In view of the difficulties associated with Janossy's theory of the cut-off, it is important to note, as remarked by George (1948), that these curves include something closely analogous to a cut-off at the low energy end. Thus (44) rises rapidly from $n(E) = 0$ when $E = m_0 c^2$ (i.e. for particles of zero momentum) to a maximum at

$$E = m_0 c^2 \sqrt{\{(5 - 2\alpha)/(3 - 2\alpha)\}},$$

which, with $\alpha = 0.2$, gives $E = 1.24 \times 10^9$ eV for protons. This appears to be in good agreement with the most recent results on the low energy spectrum referred to in § 1.

The theoretical energy flux of the arriving particles is obtained by multiplying both sides of (44) by E and integrating between appropriate limits,

$$I = \frac{\pi A' m^{(2-2\alpha)} c^{(2-4\alpha)}}{t^3} \int_{E_1}^{E_2} E \left(\frac{1 - m_0^2 c^4 / E^2}{E^{(3-2\alpha)}} \right) dE. \quad (45)$$

If comparison is to be made with the cosmic-ray data, it is reasonable to take the energy limits in the integration to be identical with the observed values. The lower limit E_1 would then be the cut-off value quoted in § 1, viz. about 2×10^9 eV; it makes little difference to the numerical value of the integral whether the upper limit E_2 is taken to be the highest energy inferred from observation (about 10^{14} eV) or the highest possible theoretical energy (infinite energy).

However, the intensity derived cannot be given an absolute value because of the undefined constant A' which appears. This constant occurs because of the arbitrary nature of the world-model adopted; before a complete comparison could be made between theory and observation the function $\psi(\xi)$ would have to be chosen more specifically.

Equating the theoretical intensity to the observed value enables an estimate to be made of A' . To obtain the intensity (total energy impinging on unit area in unit time) at the present epoch the theoretical value of the present age of the universe, reckoned in terms of the t -scale, must be substituted for t in (45); this is 2×10^9 years or 6×10^{16} sec. (WS, p. 76). If the incoming radiation is assumed to consist entirely of protons and if α is taken to be 0.2, A' proves to be of the order 10^{71} .

It is now possible to make a quantitative estimate of the particle density near the observer. However, since the approximate expression for $\psi(\xi)$ is valid only for large ξ , the total density of particles of all speeds cannot be obtained and direct comparison with observation is still impossible. The actual number of particles with speeds in the cosmic-ray velocity range is obtained by substituting appropriate values into Milne's equation (5) (WS, p. 188) and proves to be of the order 10^{-11} cm.⁻³. If the average mass of cosmic-ray particles is assumed to be about 10^{-24} g. the mass density is of the order 10^{-35} g.cm.⁻³. This is, of course, considerably less than the average density of interstellar particles of all speeds in the neighbourhood of the earth ($\sim 10^{-24}$ g.cm.⁻³). Thus the theoretical result does not appear to be inconsistent with observation, but the comparison does not provide a crucial test of the theoretical work.

10. CONCLUSIONS

By means of simplifying assumptions and a change of time-scale, it has been possible to avoid the lengthy algebra involved in Milne's original proof that free-particles accelerate to the velocity of light in finite time in his model of the universe. In the course of the investigation several general relations have been established.

By a suitable choice of fundamental observer it has been possible to treat the problem in one dimension; the equivalence of all the fundamental observers on the T -scale ensures that there is no loss of generality as a result of this procedure. Thus if a particle accelerates to the velocity of light relative to an observer in the line of sight, it must appear to do the same relative to any fundamental observer.

The analysis shows that it is unnecessary to suppose that the free-particle started its career at 'creation', as was implicitly assumed in Milne's original investigation. The particle will still reach the speed of light in finite time if it is ejected at a subsequent epoch.

The suggestion that cosmic-ray particles acquire their high speeds by a gravitational acceleration has much to recommend it. The difficulty of finding a suitable mechanism to give both the tremendous energies of individual particles and the high total intensity of the primary radiation reaching the earth is well known. The gravitational theory would appear to be capable of accounting for both: moreover it would also explain quite naturally the observed isotropy and constant intensity of the radiation (constant, that is to say, in time intervals short compared with the age of the universe). With regard to the composition of the radiation, Blackett (1936) argued against the possibility of a gravitational origin on the grounds that on any theory of this type the radiation should contain primary particles of all types, uncharged as well as charged, which he maintained was not in accord with the experimental evidence. Since then, however, heavy nuclei have been detected in the primary radiation and according to the most recent evidence it would appear that cosmic-ray particles of different types reach the earth in the same proportions as the cosmical abundances of the elements. This feature of the observations would appear not only to answer Blackett's original criticism but even to strengthen the case for a gravitational theory such as that of Milne. The energy spectrum deduced theoretically is also in good agreement with observation. No absolute value can be assigned to the total intensity to be expected, because of the arbitrary nature of the idealized world-model adopted. The density singularity at the centre of each nucleus is an imperfection in the model; whether or not the singularities are an essential feature in any model giving acceleration to the speed of light in finite time requires further investigation. Although the theory does not account for the observed preponderance of positively charged particles reaching the earth, this preponderance may well be due to some secondary mechanism and need not automatically imply an electromagnetic or magnetic primary origin.

The gravitational theory of Milne therefore provides an unforced explanation of most of the known data concerning primary cosmic radiation and is not in disagreement with the data in any respect.

It is a pleasure to express my thanks to Dr G. J. Whitrow, who first directed my attention to this problem, for many helpful discussions and suggestions.

REFERENCES

- Alfvén, H. 1937 *Ark. Mat. Astr. Fys*, **25** B, no. 29.
 Alfvén, H. 1949 *Phys. Rev.* **75**, 1732.
 Alfvén, H. 1950 *Phys. Rev.* **77**, 375.
 Babcock, H. W. 1948 *Phys. Rev.* **74**, 489.
 Blackett, P. M. S. 1936 *Halley Lecture*. Oxford University Press.
 Bradt, H. L. & Peters, B. 1950 *Phys. Rev.* **77**, 54.
 Davis, L. 1947 *Phys. Rev.* **72**, 632.
 Dolbear, D. W. N. & Elliot, H. 1950a *Proc. Phys. Soc.* **63**, 137.
 Dolbear, D. W. N. & Elliot, H. 1950b *Nature*, **165**, 353.
 Fermi, E. 1949 *Phys. Rev.* **75**, 1169.
 George, E. P. 1948 *Phil. Mag.* **39**, 240.
 Gething, P. J. D. 1950 *Phil. Mag.* **41**, 495.
 Hoang Tchang-Fong 1950 *C.R. Acad. Sci. Paris*, **230**, 292.
 Hoyle, F. 1946 *Mon. Not. R. Astr. Soc.* **106**, 384.
 Janossy, L. 1937 *Z. Phys.* **104**, 430.
 Lemaître, G. 1931 *Nature*, **128**, 704.
 Lemaître, G. 1949 *Rayons cosmiques et cosmologie*. Louvain: E. Nauwelaerts.
 Milne, E. A. 1935 *Relativity, gravitation and world structure*, Oxford University Press.
 Milne, E. A. 1948 *Kinematic relativity*, Oxford University Press.
 Pomerantz, M. A. 1950 *Phys. Rev.* **77**, 830.
 Pomerantz, M. A. & Vallarta, M. S. 1949 *Phys. Rev.* **76**, 1889.
 Richtmeyer, R. D. & Teller, E. 1949 *Phys. Rev.* **75**, 1729.
 Ryle, M. 1949 *Proc. Phys. Soc. A*, **62**, 491.
 Thiessen, G. 1949 *Observatory*, **69**, 228.
 Whitrow, G. J. 1939 *Quart. J. Math. (Oxford)*, **10**, 313.
 Whitrow, G. J. 1940 *Quart. J. Math. (Oxford)*, **11**, 53.

Oxidation-reduction processes in cultures of bacteria

I. The reducing power of *Bact. lactis aerogenes*

BY A. C. BASKETT AND SIR CYRIL HINSHELWOOD, F.R.S.

Physical Chemistry Laboratory, University of Oxford

(Received 2 May 1950)

(This paper has been printed in full in *Proceedings B*, **137**, 524)

ABSTRACT

Cultures of *Bact. lactis aerogenes* show a reducing power towards oxygen (as measured by a modification of the Thunberg methylene-blue test) which is a rather stable characteristic of the growth under specified conditions.

During adaptation of cells to utilize a given carbon substrate the reducing power and the growth rate increase in parallel (over considerable ranges) to their respective optima.

In general the reducing power of fully adapted cultures is nearly constant for a considerable variety of carbon sources (12 to 15 units) in spite of a nearly fourfold variation of optimal growth rates.

This regularity is masked with certain substrates by abnormal enhancements or inhibitions occurring at high concentrations but reappears when the measurements are made in the low concentration range.

The result suggests that there is normally a rate-limiting step in the part of the cell mechanism which consumes molecular oxygen. This is not altered by adaptation to a new carbon source, though when the cells are first transferred to a new medium the proportions of various enzymes are unsuitable for the optimum utilization of the substrate, and the amount of oxygen which can be consumed is far below the maximum. Until these proportions are finally readjusted growth rate and the degree to which the oxidizing mechanism can be used increase in parallel, a limit to the adaptive process being set by the relatively unchanging maximum oxygen uptake.

On the stability of a fluid heated from below

By O. G. SUTTON, F.R.S.

(Received 4 July 1950—Revised 13 September 1950)

In 1916 Lord Rayleigh showed that in certain circumstances a layer of fluid contained between two infinite plane horizontal surfaces could remain at rest with the density increasing upwards, and he enunciated a criterion which allows the critical temperature difference for the onset of convection to be found in terms of the depth of the fluid, the viscosity and conductivity of the gas and a reference temperature, such as the absolute temperature of the lower (heated) surface. His result was extended to a fluid bounded by two rigid plane surfaces by Jeffreys, and in this form has been examined experimentally by K. Chandra (using air) and D. T. E. Dassanayake (using carbon dioxide). Their results show (i) that the Rayleigh-Jeffreys criterion is confirmed for relatively deep layers, (ii) that in relatively shallow layers an entirely distinct mode ('columnar'), which does not satisfy the Rayleigh-Jeffreys criterion, is initiated and (iii) that this mode passes into another mode when the temperature of the lower surface is greatly increased, at temperature differences which agree approximately with Rayleigh-Jeffreys criterion.

The problem has been re-examined in the present paper, on the assumption that the new mode found by Chandra and Dassanayake arises from instability in a shallow boundary layer, whose depth is unrelated to the distance between the upper and lower rigid boundaries. It is shown that the criterion for the 'columnar' mode involves only the ratio of the absolute temperatures of the upper and lower surfaces, and an expression is derived and verified for the critical temperature difference at which the transition takes place from the 'columnar' to the 'cellular' mode. Finally, it is shown that the Rayleigh mode will occur if the depth of the fluid exceeds a certain value, but that for the more shallow layers, the 'columnar' mode will be generated initially, ultimately passing to the 'cellular' mode for increased temperature difference.

1. INTRODUCTION

The present paper is concerned with a re-examination of a problem in fluid motion which has attracted considerable attention since the initial mathematical investigation of Lord Rayleigh (1916). Physically, the problem is this: A layer of fluid, contained between two infinite plane horizontal surfaces, rigid or free, is heated from below. In certain circumstances it is found that convective motion does not begin until a critical value of the density gradient is reached, so that it is possible for a thin layer of fluid to remain permanently at rest with density increasing upwards. In this condition it must be presumed that any large differences of temperature and any tendency to sustained motion are annulled by conduction and viscosity. If the critical distribution of density is exceeded, vertical motion occurs, usually in the form of a steady tessellated pattern of ascending and descending currents, a phenomenon first described by Thomson (1882), but which is usually associated with the name of Bénard (1900, 1901), who demonstrated the formation of polygonal 'cells' in a liquid cooled from above. Because of the large diurnal changes of temperature of the earth's surface, the phenomenon is of considerable interest in meteorology, a fact first pointed out by Brunt (1939) and Low (1925).

The mathematical problem discussed by Rayleigh falls into two parts: (a) the determination of the criterion for the onset or extinction of free convection in a layer of limited depth, and (b) the investigation of the geometry of the convection pattern.

The present paper is concerned only with the first of these problems, one of the striking features of Rayleigh’s analysis being that the criterion is independent of the pattern of the cells. Rayleigh’s work has been followed by other investigations, notably those of Jeffreys (1926, 1928) and of Pellew & Southwell (1940). In all this work it is shown that the existence of a sustained convective regime depends upon the value of the non-dimensional quantity

$$Ra = -\frac{\beta g \alpha h^4}{\kappa \nu}, \tag{1}$$

which we shall call the *Rayleigh number*. Here

- β = (constant) temperature gradient between the top and bottom boundaries in initial (pre-convective) state,
- g = acceleration due to gravity,
- α = coefficient of expansion of the fluid,
- h = vertical distance between the boundaries,
- κ, ν = thermometric conductivity and kinematic viscosity, respectively, of the fluid.

The criterion derived by Rayleigh, and confirmed by other workers, is that sustained free convection is possible if, and only if,

$$\frac{\Delta \rho}{\rho(0)} > (Ra)_{crit} \frac{\kappa \nu}{gh^3}, \tag{2}$$

where $\Delta \rho$ is the difference in density of the fluid in contact with the lower and upper boundaries, $\rho(0)$ is the density of the fluid at the lower boundary and $(Ra)_{crit}$ is the critical value of the Rayleigh number appropriate to the selected boundary conditions. The principal solutions are given in table 1.

TABLE 1. CALCULATED CRITICAL VALUES OF THE RAYLEIGH NUMBER

type of boundaries	$(Ra)_{crit}$.	authority
two free-plane surfaces	657.5	Rayleigh
two rigid-plane surfaces	1707.8	Jeffreys
one free-plane surface, one rigid	1100.7	Pellew & Southwell

The implication of the criterion (2) is that as $h \rightarrow 0$, the critical density (or temperature) difference necessary for sustained convection rises without limit. This deduction has been examined by Chandra (1938) and D. T. E. Dassanayake (unpublished thesis), working at Imperial College, London, under the direction of Professor Sir David Brunt. Their results indicate a state of affairs much more complicated than that implied by (2). Chandra’s experiments were conducted with air contained between two large horizontal plane surfaces, a lower metal plate which could be heated uniformly, and an upper glass plate, through which the motion could be studied in detail. The relevant criterion for this arrangement is that of Jeffreys (1928), in which the critical value of Ra is 1708. Chandra made careful measurements of the difference of temperature between the two boundary surfaces at the stage when motion is first seen, for various depths of the chamber (h) from 4 to 16 mm. He found

that the motion set up when stability is first destroyed is not always of the same type. For 9 mm. $< h < 16$ mm. the familiar 'cellular' type was observed, with descent in the centre of the cell, for all temperature differences which exceeded the limiting value, but for shallower layers, the initial motion was of a very different type, which he called 'columnar' or 'Type II'. For $h < 6$ mm. Chandra found that motion was of the columnar type throughout the range of temperature differences employed, but for 6 mm. $< h < 9$ mm. the motion, although initially of the columnar type, changed into a cellular type when the heating of the lower plate was continued so as to increase greatly the temperature difference between the boundaries. Chandra thus clearly demonstrated the existence of two well-defined modes of free convection, and found further that the temperature differences required for the initiation of the columnar mode were small and completely different from those predicted by the Rayleigh-Jeffreys criterion. He produced some evidence, however, that the higher temperature differences required for the production of the cellular mode from the columnar mode show a measure of agreement with the Rayleigh-Jeffreys formula.

Dassanayake continued this work with carbon dioxide, for which the value of the product $\kappa\nu$ is about one-fourth of that for air, using much the same apparatus. His results, taken from the unpublished thesis, are reproduced below, by kind permission of Sir David Brunt.

TABLE 2. DIFFERENCES OF TEMPERATURE (ΔT) BETWEEN THE TOP AND BOTTOM OF A LAYER OF CARBON DIOXIDE, BOUNDED BY HORIZONTAL RIGID SURFACES, WHEN INITIAL COLUMNAR AND CELLULAR MOTION SETS IN

depth (mm.)	temp. at bottom ($^{\circ}$ C)	temp. at top ($^{\circ}$ C)	$\Delta T(^{\circ}$ C) columnar	$\Delta T(^{\circ}$ C) cellular
4	24	17½	6½	—
4½	24	16½	7½	—
4½	107	17	—	90 (approx.)
5	25	17	8	—
5	70	24	—	46
5½	26	17½	8½	—
5½	50	23	—	27
6	26	17	9	—
6	35	17½	—	17½
6½	27	17½	9½	—
6½	32	18	—	14
7	27	17	10	—
7	29	17	—	12
8	27	19	—	8
10	22	18	—	4

Dassanayake compared the results for cellular motion with the Rayleigh-Jeffreys criterion, as follows (see table 3).

The complete results of Chandra and Dassanayake and the curve for the Rayleigh-Jeffreys criterion are shown in figure 1. At least two distinct modes of motion are possible when stability is upset, one of which (cellular) agrees reasonably well with the result of the theoretical study, while the so-called 'columnar mode' shows no such agreement.

TABLE 3. EXPERIMENTAL DATA FOR THE ONSET OF CELLULAR MOTION IN A LAYER OF CARBON DIOXIDE, BOUNDED BY HORIZONTAL RIGID SURFACES, COMPARED WITH THE RAYLEIGH-JEFFREYS FORMULA

depth (mm.)	$\Delta T/T(0)$ observed	$\Delta T/T(0)$ from Rayleigh- Jeffreys formula
5	0.134	0.130
5½	0.0836	0.0904
6	0.0568	0.0623
6½	0.0459	0.0495
7	0.0397	0.0368
8	0.0267	0.0257
10	0.0115	0.0125

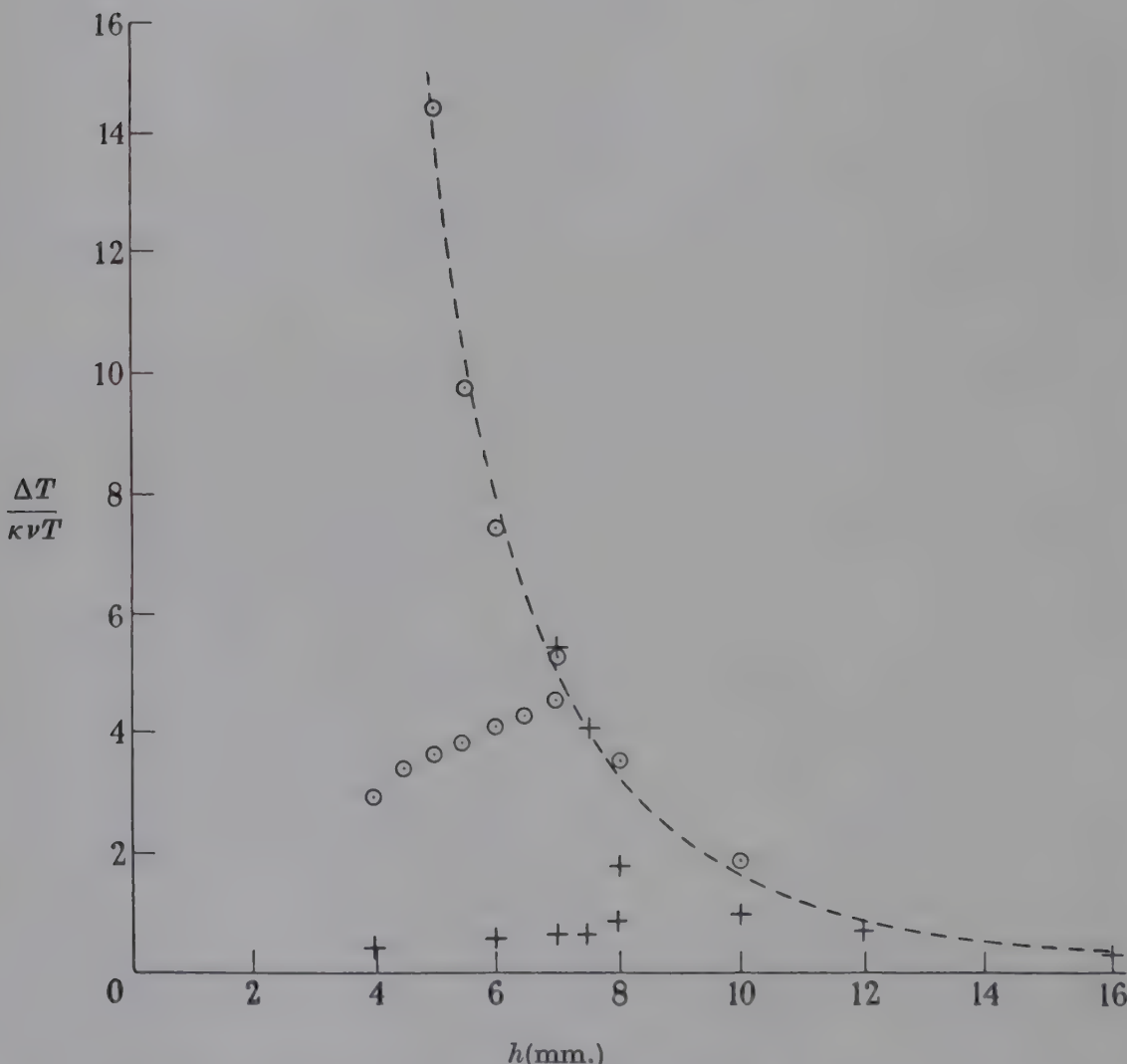


FIGURE 1. Observed values for the onset of convection between horizontal surfaces: $\odot$ observations in CO_2 ; + observations in air; --- Rayleigh-Jeffreys criterion.

2. THE RAYLEIGH ANALYSIS

In view of later discussion in this paper, the exact conditions assumed by Rayleigh and his successors are of importance, and a brief résumé of the salient points of the mathematical treatment will now be given.

It is supposed that there is an initial state of rest, in which heat is transferred by conduction in a fluid bounded only in the vertical. If the temperature is steady and uniform in all directions except in the vertical (z), the equation of conduction reduces to

$$\kappa \frac{\partial^2 T}{\partial z^2} = 0.$$

Suppose that $T_0(z)$ is the initial temperature. Then

$$T_0(z) = T(0) + \beta z, \quad (3)$$

where $T(0)$ is the temperature on the base plane $z = 0$ and β is the constant initial temperature gradient. The equation of thermal expansion is

$$\rho(z) = \rho_0(1 - \alpha\beta z), \quad (4)$$

where ρ_0 is the initial density (corresponding to T_0) and α is the coefficient of expansion. The convective motion is now supposed to begin and to be so slow that squares and products of the component velocities u, v, w (along axes x, y, z respectively) can be neglected. The corresponding deviations of temperature T' are also supposed small, so that writing

$$T = T_0 + T' = T(0) + \beta z + T', \quad (5)$$

all second-order terms in u, v, w and T' are negligible. The modified equation of conduction is then

$$\frac{\partial T'}{\partial t} + w\beta = \kappa \nabla^2 T'. \quad (6)$$

The problem now becomes one of incompressible viscous motion, the Navier-Stokes equations reducing to

$$\left. \begin{aligned} \frac{\partial u}{\partial t} &= -\frac{1}{\rho} \frac{\partial p'}{\partial x} + \nu \nabla^2 u, \\ \frac{\partial v}{\partial t} &= -\frac{1}{\rho} \frac{\partial p'}{\partial y} + \nu \nabla^2 v, \\ \frac{\partial w}{\partial t} &= -\frac{1}{\rho} \frac{\partial p'}{\partial z} + \nu \nabla^2 w + g\alpha T', \end{aligned} \right\} \quad (7)$$

where p' is the deviation of pressure from that given by the hydrostatic equation and ρ is regarded as a constant. The assumption of incompressibility is justified by the fact that the change in ρ brought about by the motion is a fraction $-\alpha T'/(1 - \alpha\beta z)$ of the steady value, and is thus negligible when multiplied by u, v, w, T' or p' . Hence the expansion of the fluid by heating has no dynamical consequences except in the one equation (that for w) where it appears in the external force, that is, when α is multiplied by g .*

The result of eliminating u, v, p' and T' from equations (6) and (7) and the equation of continuity is a sixth-order partial differential equation in w :

$$\left(\frac{\partial}{\partial t} - \kappa \nabla^2\right) \left(\frac{\partial}{\partial t} - \nu \nabla^2\right) \nabla^2 w + \beta \alpha g \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right) = 0. \quad (8)$$

* Rayleigh (1916) expresses this by the somewhat cryptic phrase that the change of density with temperature is immaterial 'except in so far as it modifies the operation of *gravity*'.

The boundary conditions are:

$$(1) \text{ at a rigid surface } w = \frac{\partial w}{\partial z} = 0,$$

$$(2) \text{ at a free surface } w = \frac{\partial^2 w}{\partial z^2} = 0,$$

$$(3) T' = 0 \text{ on both surfaces.}$$

In equation (8) write

$$w = f(x, y) F(z) e^{\sigma t}.$$

In this expression $f(x, y)$ specifies the type of tessellation; exact expressions for $f(x, y)$ are known for square and hexagonal cells (Rayleigh 1916; Christopherson 1940), but are not necessary for the determination of the criterion of stability. For sustained convective motion it may be shown that the real part of σ must be positive (Pellew & Southwell 1940) and convection will just begin when $\sigma = 0$, in which case all time variations are zero, the problem becoming one of steady motion. Equation (8) becomes

$$\nabla^6 w = -\frac{\beta \alpha g}{\kappa \nu} \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} \right).$$

The depth of the layer (h) is now introduced by the change of variable $\zeta = z/h$, which transforms the equation into one in which the constants are grouped in the non-dimensional combination $-\beta \alpha g h^4 / \kappa \nu$, which we have called the Rayleigh number. The determination of the criterion for the onset of convection thus amounts to the purely numerical problem of finding the least value of Ra for which steady motion can take place. The principal results have been given above in table 1.

To obtain a simple physical interpretation of these results Rayleigh puts

$$\beta = \Delta T / h, \quad \Delta T = T(0) - T(h),$$

$$\alpha = \Delta \rho / (\rho(0) \Delta T), \quad \Delta \rho = \rho(h) - \rho(0),$$

and the criterion in the form (2) follows immediately.

If h is small and derivatives of ρ higher than the first are neglected

$$\rho(h) = \rho(0) + h \left(\frac{d\rho}{dz} \right),$$

or

$$\frac{\Delta \rho}{\rho(0)} = \frac{h}{\rho(0)} \left(\frac{d\rho}{dz} \right)_0.$$

For a perfect gas at rest in a gravitational field

$$\frac{1}{\rho} \frac{d\rho}{dz} = -\frac{1}{T} \left(\frac{dT}{dz} + \frac{g}{R} \right).$$

The quantity g/R , (R = gas constant) is the so-called 'auto-convection' gradient, whose value for dry air is about $3.5 \times 10^{-4} \text{ } ^\circ\text{C cm}^{-1}$. The criterion may thus be expressed in the form

$$-\left(\frac{dT}{dz} + \frac{g}{R} \right)_0 > (Ra)_{\text{crit}} \frac{\kappa \nu}{g h^3} \frac{T(0)}{h}, \quad (10)$$

or, in words, that convection will occur when the difference between the observed gradient and the auto-convection gradient attains a certain value. Since the gradient of temperature is assumed constant in the initial state, (10) may be replaced by

$$\frac{\Delta T}{T(0)} > (Ra)_{\text{crit.}} \frac{\kappa \nu}{gh^3} - \frac{hg}{RT(0)}.$$

With gases such as air or carbon dioxide, if h is of the order of 1 cm. or less, the term $hg/RT(0)$ may be neglected when $T(0)$ is of the order of 300°K , that is, at normal (absolute) terrestrial temperatures. The criterion in the form appropriate to the experiments of Chandra and Dassanayake is thus

$$\frac{\Delta T}{T(0)} > (Ra)_{\text{crit.}} \frac{\kappa \nu}{gh^3} = \frac{1708\kappa \nu}{gh^3}. \quad (11)$$

3. CRITICAL EXAMINATION OF THE RAYLEIGH ANALYSIS

Rayleigh's treatment contains a number of important assumptions and simplifications, these being principally:

(1) The neglect of all density variations except that which is directly associated with buoyancy.

(2) The assumption that the initial motion is slow.

(3) The assumption that at the stage of 'just-no-convection', a linear temperature profile extends throughout the entire layer of the fluid.

Of these, (1) and (2) may be accepted as reasonable, but (3) raises difficulties. Ramdas & Malurkar (1932) have made observations of temperature very near a heated surface, and their results indicate that in steady free convection there exists a shallow 'surface layer', in which temperature falls very rapidly and uniformly with height (as much as 1 or 2°C per mm.) together with an overlying region, of indefinite height, within which the temperature gradient is much smaller. As far as can be judged from the curve given by these authors (1932, figure 4), the region of approximately constant temperature gradient is about 1 cm. or less in depth. These observations refer to a fluid which is unbounded above. Chandra (1938, p. 235) in describing columnar motion states that 'After a brief interval [following the introduction of smoke into the chamber] a shallow layer of clear air forms immediately above the heated plate, this being the dust-free space discussed by Aitken, and the clear air bursts upwards to the top of the chamber in small vertical columns'.

These observations throw doubt on the fundamental concept of Rayleigh's method, that at the stage of 'just-no-convection' a linear temperature profile necessarily exists throughout the whole depth of the chamber. It appears more likely that the process of the initiation of convection is as follows: When heating of the lower surface begins, a large, approximately constant, gradient of temperature is created in a shallow surface layer. As heating proceeds, this layer becomes unstable and breaks up to form the columnar vortices described by Chandra. Since one of the fundamental assumptions of Rayleigh's work is not satisfied, it is not surprising that the temperature difference over the whole depth of the chamber at this stage differs from that given by the theoretical criterion. As heating of the lower plate proceeds,

however, it is conceivable that the conditions envisaged by Rayleigh are approached; a linear temperature profile ultimately extends from the bottom surface to the top. At this stage cellular motion sets in, and the critical value of the temperature difference should be in good agreement with that derived mathematically. Rayleigh (1916) appears to have had some doubts regarding the assumption of a linear fall of temperature extending throughout the entire layer, for he remarks: 'In an experiment the temperature gradient could not be established all at once and we may suppose the process to be very slow. In earlier stages the equilibrium would be stable (no convection: conduction only) so that no disturbance of importance would occur until...the thermal gradient β reached the critical value.'

Rayleigh's formulation of the problem thus appears to correspond to an experimental arrangement in which the temperature of the lower surface is raised very slowly, so that an accumulation of heat near the surface and the formation of a distinct boundary layer with a large temperature gradient is unlikely. It is improbable that this condition was strictly satisfied in the experimental work, and we therefore proceed to investigate the problem by assuming that the observed initial motion arises from instability in a thin layer whose depth is not necessarily related to the arbitrary distance between the upper and lower boundaries chosen by the experimenter.

4. DIMENSIONAL ANALYSIS

Consider the following idealized form of the experiments of Chandra and Dassanayake: A layer of gas, initially at rest and of uniform temperature, is contained between two infinite horizontal plane surfaces, the upper of which is kept at a fixed temperature. If the rate of heating of the lower surface is sufficiently rapid, there will be an accumulation of heat and consequently a steep temperature gradient in a thin layer near the surface. If the layer becomes unstable, the resulting convection will reduce the gradient of temperature near the surface, but if the rate of heating is insufficient, the motion will decay. We assume that at the stage of 'just-no-convection', the loss of heat from the lower surface by convection is equal to that by conduction through a certain layer, of depth δ , and we proceed to find an expression for δ .

Since the distribution of temperature is linear in δ ,

$$\frac{\Delta T}{\delta} = -\left(\frac{dT}{dz}\right)_0, \quad (12)$$

where $\Delta T = T(0) - T(\delta)$. The loss of heat by conduction from the lower surface is

$$H = -\kappa c_p \rho \left(\frac{dT}{dz}\right)_0 = -\kappa c_p \rho \frac{\Delta T}{\delta}. \quad (13)$$

In a steady state of convection, it is well established that the heat loss from a horizontal plane surface depends primarily on $(\Delta T)^{\frac{1}{4}}$, where ΔT is the difference in temperature between the surface and the main body of the fluid, well above the surface layer. Fishenden & Saunders (1932) give the empirical equation

$$H = 6 \times 10^{-5} (\Delta T)^{\frac{1}{4}}, \quad (14)$$

where H is expressed in $\text{g.cal.cm.}^{-2}\text{sec.}^{-1}$ and ΔT in $^{\circ}\text{C}$. This appears to be valid over a wide range of surface temperatures. From equations (13) and (14),

$$\delta = \frac{\kappa c_p \rho}{6 \times 15^{-5}} (\Delta T)^{-\frac{1}{2}}. \quad (15)$$

The depth of the molecular conduction layer, being unrelated to the distance between the upper and lower boundaries, must involve some scale of length inherent in the properties of the fluid. Such a scale is provided by $(\kappa^2/g)^{\frac{1}{2}}$ or by $(\kappa\nu/g)^{\frac{1}{2}}$ in the present instance, in which both conduction and viscosity are involved. Thus the depth δ may be written tentatively in the form

$$\delta = \text{constant } (\kappa\nu/g)^{\frac{1}{2}} (\alpha\Delta T)^{-\frac{1}{2}}, \quad (16)$$

where α , the coefficient of expansion of the fluid, is here taken to be $1/T(h)$, that is, referred to the fixed temperature of the upper surface as a reference temperature.

If the stability of the layer δ be examined by Rayleigh's method, equation (9), with the differential coefficients replaced by finite differences, will involve only the temperatures $T(0)$ and $T(h)$ and their difference, since κ and ν disappear because $\delta^3 \propto (\kappa\nu/g) (\alpha\Delta T)^{\frac{1}{2}}$. If variations in κ and ν are neglected, the criterion for the onset of convection starting in the δ -layer should thus involve only the ratio $T(0)/T(h)$. The transition to the cellular mode implies entirely different considerations, since in this case the δ -layer may be expected to coincide with the whole layer of fluid. The analysis is not conclusive, but suggests the desirability of an empirical study of the experimental data on the above lines.

5. EMPIRICAL STUDY OF THE OBSERVATIONS

We note first that equation (15) yields results of the correct order of magnitude. For $\kappa = 0.2 \text{ cm.}^2\text{sec.}^{-1}$, $c_p = 0.25$, $\rho = 1.2 \times 10^{-3} \text{ g.cm.}^{-3}$, values appropriate to air at about 27°C , we have

$$\delta = \frac{6.25 \times 10^{-5}}{6 \times 10^{-5}} (\Delta T)^{-\frac{1}{2}} \simeq (\Delta T)^{-\frac{1}{2}}.$$

For $1^{\circ} \leq \Delta T \leq 25^{\circ}$ this implies $0.4 \text{ cm.} \leq \delta \leq 1 \text{ cm.}$ Thus the values of δ agree in magnitude with the depth of the 'linear profile' layer found by Ramdas & Malurkar (1932).

(a) Columnar mode

The observations in carbon dioxide and in air for this type of motion are found to satisfy the empirical equation

$$\frac{\Delta T}{T(0)} = 0.507 \left(\frac{\Delta T}{T(h)} \right)^{\frac{1}{2}} - 0.0074 \quad (17)$$

to a high degree of accuracy. This is shown in table 4.

This relation bears out the expectation of the theoretical arguments of the preceding section, that for this type of convection the condition for steady motion involves only the absolute temperatures of the boundaries and is independent of the depth of the chamber and the molecular constants of the gas.

TABLE 4. DATA FOR THE ONSET OF COLUMNAR CONVECTION IN CO₂ AND IN AIR, COMPARED WITH THE EQUATION

$$\Delta T/T(0) = 0.507(\Delta T/T(h))^{\frac{2}{3}} - 0.0074$$

gas	T(0)	T(h)	(1)	(2)	difference (1 - 2)	chamber depth (mm.)
			$\Delta T/T(0)$	$0.507(\Delta T/T(h))^{\frac{2}{3}} - 0.0074$		
CO ₂	297	290.5	0.0219	0.0220	- 0.0001	4
	297	289.5	0.0253	0.0254	- 0.0001	4½
	298	290	0.0269	0.0269	negligible	5
	299	290.5	0.0284	0.0284	negligible	5½
	299	290	0.0301	0.0301	negligible	6
	300	290.5	0.0317	0.0316	+ 0.0001	6½
	300	290	0.0333	0.0332	+ 0.0001	7
	306.8	299	0.0254	0.0255	- 0.0001	4
air	307.3	300.3	0.0228	0.0228	negligible	6
	305.3	297.8	0.0246	0.0247	+ 0.0001	7
	305.3	300	0.0174	0.0172	+ 0.0002	7½

From equation (17) it follows that $x = T(0)/T(h)$ satisfies the equation

$$1 - \frac{1}{x} = 0.507(x - 1)^{\frac{2}{3}} - 0.0074, \tag{18}$$

at the onset of columnar motion. This equation has two real roots, $x = 1.021$ (approx.) and $x = 1.031$ (approx.), but any value of x between the limits 1.02 and 1.035 may be regarded as an approximate root within the limits of experimental error, since the curves $y = x - 1/x$ and $y = 0.507(x - 1)^{\frac{2}{3}} - 0.0074$ are almost identical over this range of values of x . The deduction of the empirical study is therefore that columnar motion will be initiated when the ratio $T(0)/T(h)$ lies between 1.02 and 1.035. That this is so is shown by the following data:

gas	CO ₂	CO ₂	CO ₂	CO ₂	CO ₂	CO ₂	CO ₂	air	air	air	air
observed $T(0)/T(h)$	1.022	1.026	1.028	1.029	1.031	1.033	1.034	1.026	1.023	1.025	1.018
chamber depth (mm.)	4	4½	5	5½	6	6½	7	4	6	7	7½

The experimental data are not entirely consistent in that the ratio $T(0)/T(h)$ for CO₂ shows a small but steady increase with increasing h , whereas the ratio for air shows no systematic relation with h . The most striking fact is that the ratios found are the same for both gases, despite the great difference (roughly 4 : 1) between the values of $\kappa\nu$ for air and CO₂.

If $T(0)/T(h)$ falls outside the range 1.02 to 1.035 it should be impossible to determine a definite value of ΔT for the onset of columnar motion. This again is borne out by the experiments. For CO₂, the upper surface was maintained at about 290° K, so that for columnar motion $T(0)$ should lie between 296 and 300° K. Dassanayake's range for his experiments was 297 to 300° K. For air, with $T(h) = 300^\circ$ K, the theoretical range for $T(0)$ is 306 to 310° K. Chandra used the range 305.3 to 307.3° K, and records that a definite end-point could not be found with a very shallow chamber and $T(0) = 301.3^\circ$ K.

(b) Cellular motion

Both Chandra and Dassanayake found that, provided the depth of the chamber lies between certain limits, columnar motion gives place to cellular motion when the temperature of the lower surface is greatly increased. From the physical theory of §3, cellular motion is initiated when a linear temperature profile extends over the whole chamber, so that effectively $\delta = h$. From equation (16) this implies that the expression

$$h(\Delta T/T(h))^{\frac{1}{2}}(g/\kappa\nu)^{\frac{1}{2}}$$

should be an absolute constant (K). This deduction can be tested with the experimental data for CO_2 , but for air a difficulty arises in that Chandra (1938) does not give the upper surface temperature for the transition cases, but only the temperature differences. It has therefore been assumed that $T(h) = 300^\circ \text{K}$ for these experiments, this being the temperature of the water bath employed for the upper surface in the columnar mode determinations.

TABLE 5. VALUES OF $h(\Delta T/T(h))^{\frac{1}{2}}(g/\kappa\nu)^{\frac{1}{2}} = K$ FOR THE TRANSITION FROM COLUMNAR TO CELLULAR MOTION

gas	depth (h) of layer (mm.)	ΔT ($^\circ\text{C}$)	$T(h)$ ($^\circ\text{K}$)	$\kappa\nu$ cm. ⁴ sec. ⁻²	K
CO_2	$4\frac{1}{2}$	90 (approx.)	290	0.0115	14.8
	5	46	297	0.010	14.6
	$5\frac{1}{2}$	27	296	0.009	14.5
	6	$17\frac{1}{2}$	290	0.0079	14.9
	$6\frac{1}{2}$	14	291	0.0078	15.3
air	7	103	(300)	0.048	14.6
	$7\frac{1}{2}$	62	(300)	0.044	14.6
	8	23	(300)	0.035	12.8

The deduction is well supported by the data (table 5). For this range of values of h the critical temperature difference is given by

$$\Delta T = T(h) K^4 \left(\frac{\kappa\nu}{gh^3} \right)^{\frac{1}{2}} = T(h) K^4 \left(\frac{\kappa\nu}{g} \right)^{\frac{1}{2}} \frac{1}{h^{\frac{3}{2}}}. \quad (19)$$

Omitting the observation in air for $h = 8$ mm., the mean value of K is 14.7. Table 6 shows the comparison between observation, the Rayleigh-Jeffreys formula and the expression (19). These results show that the expression (19) is in slightly better agreement with the observations for shallow layers and large temperature differences than the Rayleigh-Jeffreys formula. On the other hand, the Rayleigh-Jeffreys criterion gives results much closer to the observed values when no transition takes place, that is, when h is relatively large.

These results indicate a possible explanation for one of the most striking features of the experimental work. Both experimenters found that if h exceeds a certain value (depending on the nature of the gas), the Rayleigh mode is generated *ab initio* and is apparently permanent. If, however, h is less than these critical values, the initial mode is the columnar mode, the lower limit for which has been shown to be $\Delta T/T(0) = 0.021$. Thus the Rayleigh mode should appear as the initial mode if h is

such that the Rayleigh-Jeffreys criterion indicates values of $\Delta T/T(0) < 0.021$. The condition for this is that h must exceed a value h_c given by

$$\frac{1708\kappa\nu}{gh_c^3} = 0.021.$$

For air, taking $\kappa\nu = 0.03 \text{ cm}^4 \text{ sec}^{-2}$, this gives

$$h_c = 13\frac{1}{2} \text{ mm. (approx.)}.$$

For CO_2 , taking $\kappa\nu = 0.0075 \text{ cm}^4 \text{ sec}^{-2}$, the corresponding depth is

$$h_c = 8\frac{1}{2} \text{ mm. (approx.)}.$$

Chandra's results (1938, table 1, p. 233) show that reasonably close agreement with the Rayleigh-Jeffreys criterion was obtained for $h \geq 12 \text{ mm.}$, and Dassanayake's data (table 2 of the present paper) indicate that the corresponding result for CO_2 is $h \geq 8 \text{ mm.}$ The agreement is fairly satisfactory.

TABLE 6. OBSERVED VALUES OF ΔT FOR THE TRANSITION FROM COLUMNAR TO CELLULAR MOTION, COMPARED WITH THEORY

gas	depth (mm.)	ΔT observed (°C)	ΔT from Rayleigh- Jeffreys criterion (°C)	ΔT from equation (19) (°C)
CO_2	$4\frac{1}{2}$	90 (approx.)	73	88
	5	46	44.7	47.2
	$5\frac{1}{2}$	27	29.2	29.0
	6	$17\frac{1}{2}$	19.2	16.9
	$6\frac{1}{2}$	14	15.1	12.2
air	7	103	84.2	104
	$7\frac{1}{2}$	62	58.7	59.6
	8	23	36.3	47.5

6. CONCLUSIONS

The results of this investigation suggest the following description of natural convection in a layer of gas bounded by large horizontal surfaces, the upper of which is maintained at constant temperature. If the layer is deep, Rayleigh's analysis shows that motion will be initiated for small values of ΔT . For these depths, the Rayleigh-Jeffreys analysis appears to be substantially confirmed, and no transition to any other type of steady motion seems to be possible once this mode is established. For smaller depths, the analysis shows that the so-called columnar mode is initiated for values of ΔT much below those indicated by the Rayleigh-Jeffreys criterion. The condition for this type of motion is that $T(0)/T(h)$ must satisfy equation (19), or that $\Delta T/T(0)$ is approximately constant. Since the acceleration due to buoyancy is $g\alpha\Delta T$, which depends only on the ratio $T(0)/T(h)$ in a layer possessing a linear temperature profile, columnar convection is consistent with instability in a boundary layer whose depth depends, *inter alia*, on the nature of the gas, so that this type of motion requires that the acceleration experienced by a volume of the gas shall attain a certain value, independent of the nature of the gas or the depth of the chamber. If the temperature of the lower surface is greatly increased, the motion changes to

one which in appearance approximates to the Rayleigh type, the condition for transition being given by equation (19).

The writer desires to express his thanks to Professor Sir David Brunt, Sec.R.S., who kindly made available the original theses by K. Chandra and D. T. E. Dassanayake, and to his colleagues of the Physics Branch of the Military College of Science for much helpful discussion.

REFERENCES

- Bénard, H. 1900 *Rev. Gen. Sci. Pur. Appl.* **12m** 1261, 1309.
 Bénard, H. 1901 *Liebigs Ann.* **23**, 62.
 Brunt, D. 1939 *Physical and dynamical meteorology*, 2nd ed. p. 219. Cambridge University Press.
 Chandra, K. 1938 *Proc. Roy. Soc. A*, **164**, 231.
 Christopherson, D. G. 1940 *Quart. J. Math.* **11**, 63.
 Fishenden, M. & Saunders, O. A. 1932 *The calculation of heat transmission*. London: H.M.S.O.
 Jeffreys, H. 1926 *Phil. Mag.* **2**, 833.
 Jeffreys, H. 1928 *Proc. Roy. Soc. A*, **118**, 195.
 Low, A. R. 1925 *Nature*, **115**, 299.
 Pellew, Anne & Southwell, R. V. 1940 *Proc. Roy. Soc. A*, **176**, 312.
 Ramdas, L. A. & Malurkar, S. L. 1932 *Indian J. Phys.* **7**, 1.
 Rayleigh, Lord 1916 *Phil. Mag.* **32**, 529.
 Thomson, J. 1882 *Proc. Glasg. Phil. Soc.* **13**, 464.

An experimental and thermodynamic investigation of the hydrogen-titanium system

By A. D. McQUILLAN, *Baillieu Laboratory, University of Melbourne, Australia*

(Communicated by Sir David Rivett, F.R.S.—Received 18 May 1950—
Revised 27 July 1950)

The temperature-pressure-concentration relationships of the hydrogen-titanium system have been studied, and its more important thermodynamic aspects considered. Part of the equilibrium diagram has been established, and free-energy curves for the α - and β -solutions have been plotted for a number of temperatures. The experimental results have been used to obtain values for the latent heat of the $\alpha \rightleftharpoons \beta$ transformation and the transformation temperature.

INTRODUCTION

When titanium, at any given temperature above about 400°C, is placed in contact with hydrogen, the gas is taken up until the concentration of hydrogen in the metal reaches a particular value which is dependent only on the pressure of the external hydrogen. The absorption of the hydrogen by titanium is, therefore, a reversible process which can be treated thermodynamically. In this work, the equilibrium relationships between temperature, pressure and concentration have been studied,

with the object of establishing the equilibrium diagram of the hydrogen-titanium system, and obtaining as much thermodynamic information relating to the system as possible.

Earlier work on the hydrogen-titanium system suffered from the fact that only very impure titanium appeared to have been available. Hägg (1931) studied the crystal structure of a range of hydrogen-titanium alloys, but found that the lattice parameter of one of the phases in a two-phase region appeared to increase with increasing hydrogen concentration. Kirschfeld & Sieverts (1930) determined the isothermal pressure-concentration curves at four different temperatures. These results have been used by Barrer (1944) to calculate various thermodynamic functions for the system. With the purer material now available, appreciable differences from the results of these early workers have been observed.

As in the case of any other equilibrium diagram, the study of the hydrogen-titanium system was limited at about 400°C by the increasingly long periods required for the attainment of equilibrium. The investigation of the high-temperature region was also limited in this instance by the lack of a refractory material with which titanium will not react, and which will remain sufficiently impervious to gases at temperatures above 950°C, the maximum temperature at which silica has been found to be satisfactory. A further limitation was imposed by the use of glass vacuum apparatus, in which pressures above 1 atm. could not be contained. In spite of these restrictions, a sufficiently large region of the constitutional diagram has been covered to yield useful and interesting results.

The temperature region investigated includes the transformation of pure titanium, which occurs at about 880°C. The effect of hydrogen on this transformation has been determined, and the existence of a third phase, γ , at higher hydrogen concentrations has been established.

It has been found possible to formulate equations for the temperature-pressure-concentration relationships in the α - and β -regions, and from them to deduce the changes in free energy of these two phases with hydrogen concentration at a number of temperatures. The difference between the values of the free energies of the two forms of the pure metal has been determined as a function of temperature. A value for the transformation temperature and the latent heat of transformation of the pure metal has been calculated from the chemical activities of the titanium in the α - and β -solutions.

EXPERIMENTAL

For experimental simplicity the changes of pressure occurring with temperature when titanium metal was heated in a closed system in contact with a predetermined quantity of hydrogen were measured. Since, however, the volume of the system must be finite, the amount of hydrogen outside the metal required to set up the pressure is also finite, and increases with the pressure set up. The total amount of hydrogen in the system being constant, any increase in the quantity of external hydrogen involves a decrease in the hydrogen concentration in the metal, which thus changes throughout the experiment. Since the volume of the apparatus is known, the concentration of hydrogen in the metal may be calculated for any pressure. Thus, for

any point on the experimental curves, pressure, temperature and concentration are known, and hence the relationships between pressure and temperature at constant concentration, and between pressure and concentration at constant temperature may be derived.

Apparatus

In order to ensure, without increasing the size of the specimen unduly, that the amount of hydrogen outside the titanium specimen was small compared with that contained in it, the internal volume of the experimental apparatus was reduced to a minimum, capillary tubing being used wherever possible. The titanium specimen itself was contained in a clear silica tube 5 mm. in diameter. The volume of hydrogen at a temperature higher than that of the room was reduced to negligible proportions by filling the part of the heated specimen tube not occupied by the titanium with a closely fitting silica rod.

The pressures measured during the course of the work ranged from 0.1 to 760 mm. of mercury. From 0.1 to 20 mm. a dibutyl phthalate manometer was used. Above 20 mm. it was replaced by a mercury manometer. The vapour pressure of dibutyl phthalate at room temperature is negligibly small; its density is one-thirteenth that of mercury, and it has the further advantage that it does not stick in the manometer tube at low pressures. To prevent bubbles of air from forming in the liquid during its introduction into the manometer, the dibutyl phthalate was distilled into the degassed tube under vacuum.

In work of this kind it is necessary to ensure that the apparatus is as free from contaminating gases as possible. It was not possible to bake out the complete apparatus, but as much of it as possible was heated under vacuum and the whole system continuously evacuated at a pressure of 10^{-6} mm. of mercury for several days. The grease used on the taps was vacuum-distilled to remove all traces of dissolved air. It was found that, when isolated from the pumps, the apparatus would hold a vacuum of 10^{-5} mm. of mercury for several days. As a final check on its freedom from contaminating gases, a piece of pure titanium was heated in the evacuated and sealed apparatus for 54 hr. at 1000°C . No increase in the hardness of the metal resulted, indicating that no serious contamination had occurred. The affinity of titanium for oxygen and nitrogen is so great, however, that it was expected that some slight pick-up of these gases might occur during a series of experimental runs owing to the manipulations involved. It is known that the diffusion rate of the gases is slow, even at 950°C , and it was considered that during the time of an experiment, penetration of the metal by oxygen and nitrogen would not have progressed much past the surface layers, leaving the inner part oxygen- and nitrogen-free. For this reason a block specimen was used rather than powder. The time required for the attainment of equilibrium with respect to the hydrogen was not appreciably affected by the form of the specimen. The specimen was arranged so that it touched the silica wall of the specimen tube at only two points, thereby reducing the chances of contamination from this source.

The titanium was prepared by Philips Gloeilampenfabrieken, Eindhoven, using the method originally developed by van Arkel, and described by Fast (1939), in which the iodides of titanium are decomposed on a hot filament. The metal was very

ductile, and had a hardness of 70 v.p.h. Spectrographic analysis showed it to contain small quantities of silicon, iron, chromium, magnesium, antimony and copper, the total of these impurities amounting to less than 0.07 %. Pure hydrogen was produced by heating titanium hydride. The gas was held at a predetermined pressure in a reservoir of accurately known volume, from which it was admitted to the part of the apparatus containing the titanium specimen, the pressure in the reservoir before and after admission of the hydrogen giving a measure of the amount put in. The volume of this part of the apparatus itself was determined by measuring the pressure set up in it by a known quantity of hydrogen. Since slight changes of volume occur as the pressure in the apparatus changes, owing to the movement of the manometer liquid, it was determined for the whole range of pressures encountered. The temperature of the furnace in which the specimen tube was heated was controlled to within $\pm 0.5^\circ\text{C}$ of any required temperature.

The apparatus used is shown in figure 1.

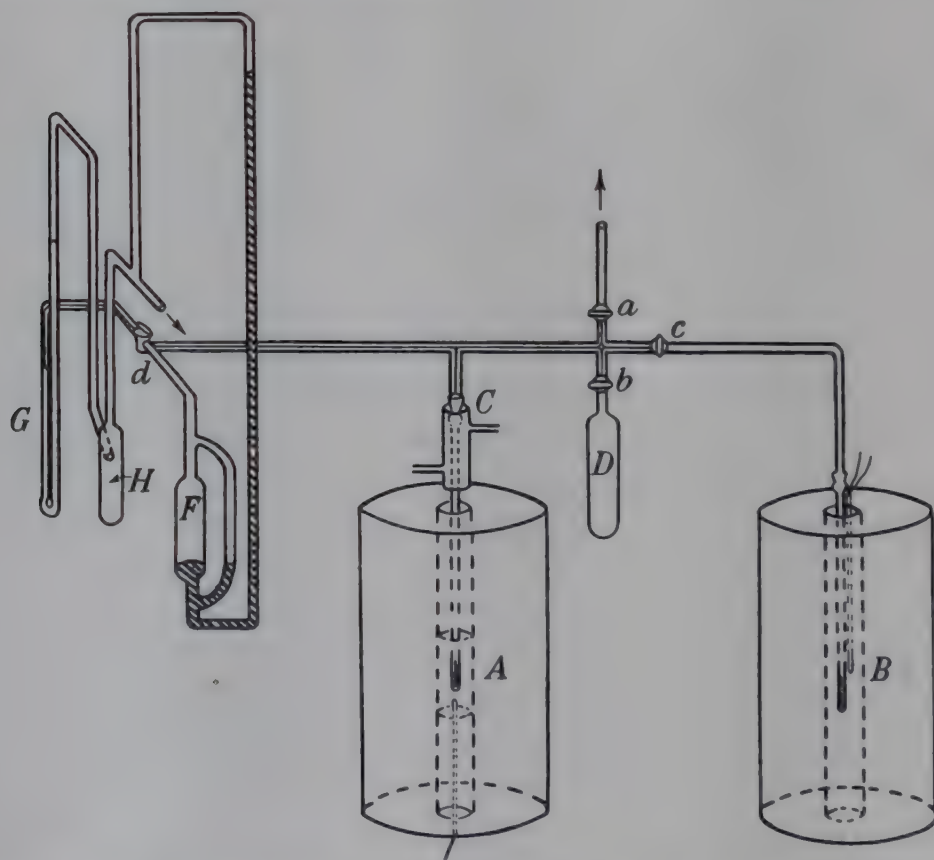


FIGURE 1. Apparatus. The specimen, contained in a silica tube, is heated in furnace *A*. *C* is a water-cooled vacuum joint by which the specimen tube is connected to the remainder of the apparatus. Furnace *B* contains titanium hydride and provides the hydrogen necessary for the experiments. *D* is a gas reservoir of accurately known volume, while *F* and *G* are, respectively, mercury and dibutyl phthalate manometers. *H* is a liquid trap to prevent the manometric liquids from being sucked into the pumping system. Vacuum taps, *a*, *b*, *c*, and the two-way tap *d* are used to isolate various parts of apparatus. Connexions to pumps are shown by arrows.

Procedure

Although theoretically the complete investigation could be made using a single specimen of titanium, in practice no specimen was used with more than four different

quantities of hydrogen, since both the error in the estimation of the amount of hydrogen present involved in introducing successive amounts into the apparatus, and the contamination of the specimen by oxygen and nitrogen are cumulative.

Pressures were measured at 10°C intervals over the temperature range between that at which the time of attainment of equilibrium was impracticably long, and that at which the pressure equalled one atmosphere. Above 500°C the equilibrium pressure was attained almost instantaneously as long as the hydrogen-titanium alloy remained a homogeneous single phase. In regions in which two metallic phases co-existed, the time required for the attainment of equilibrium was much longer, often as long as 1 hr., owing to the fact that nucleation and growth of the second phase had to occur. Identical curves for pressure against temperature were obtained both on heating and cooling.

THE PRESSURE-TEMPERATURE-CONCENTRATION CURVES

The full lines in figures 2 and 3 show the curves of pressure against concentration at constant temperature and of pressure against temperature at constant concentration. Concentration and pressure have been plotted logarithmically and temperature reciprocally.

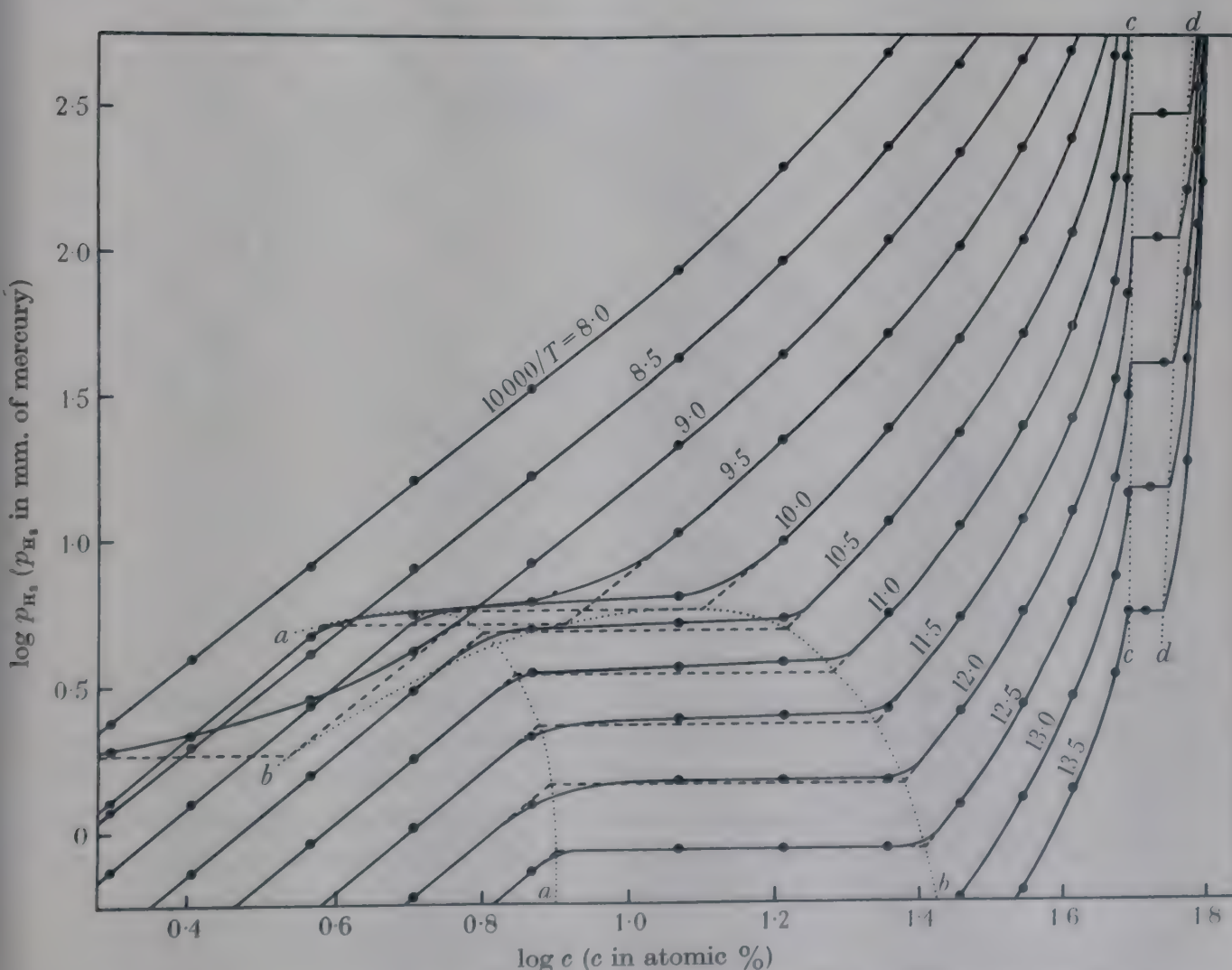


FIGURE 2. Isothermal pressure-concentration curves for the hydrogen-titanium system.

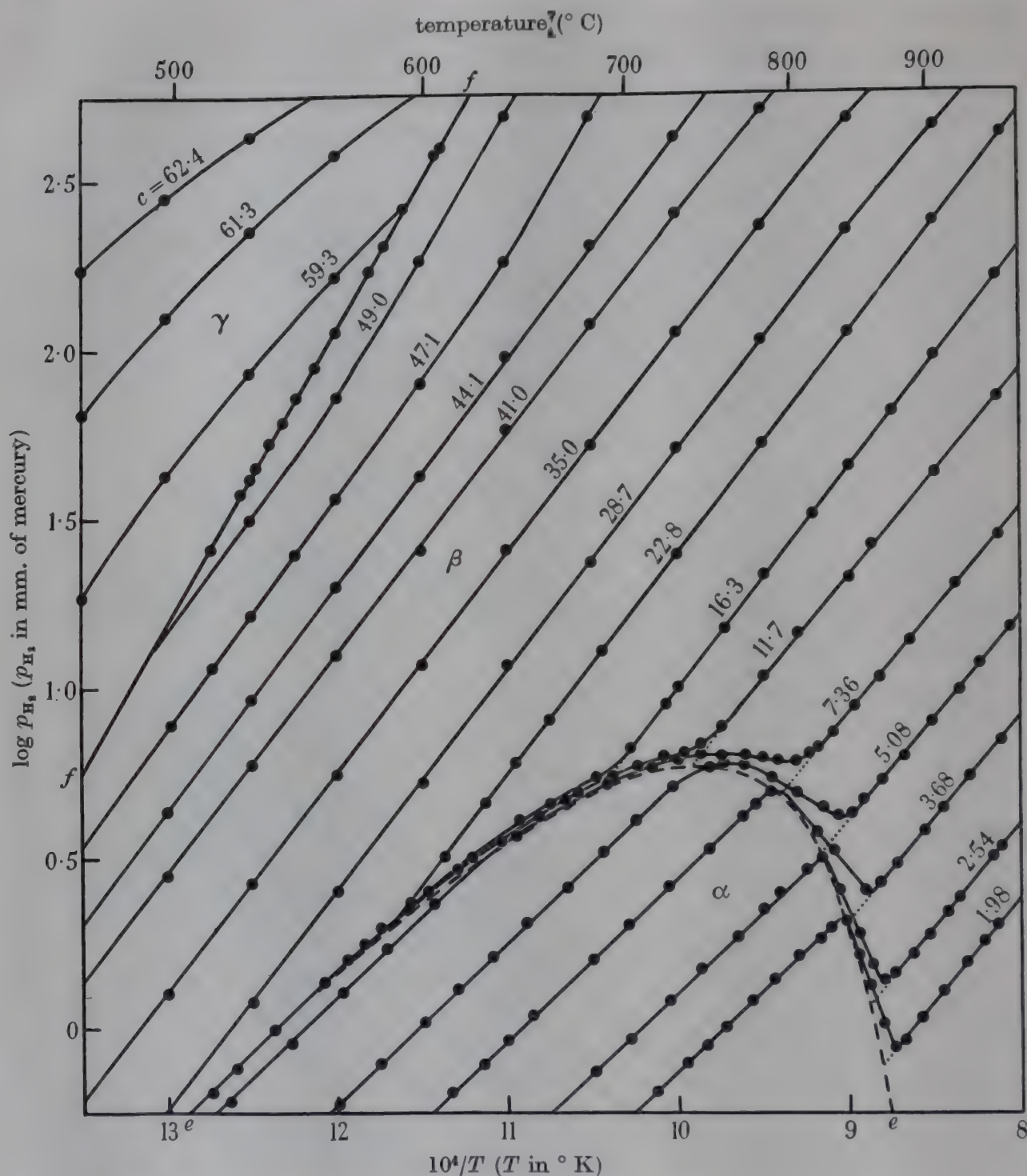


FIGURE 3. Pressure-temperature curves for range of concentrations in the hydrogen-titanium system.

It follows directly from the phase rule that in single-phase regions of the equilibrium diagram, the pressure of the hydrogen in equilibrium with the titanium-hydrogen alloy at constant temperature must vary with the concentration of hydrogen in the alloy, whereas in two-phase regions the pressure must be constant. The terms single- and two-phase region mentioned above refer to the condensed system and will be used in that sense throughout the following discussion of the results. In figure 2 it can be seen that the pressure first increases as the concentration of hydrogen is increased. The pressure-concentration curve then turns sharply, becoming parallel or nearly parallel to the concentration axis, indicating a two-phase region. The curve

subsequently continues upwards, until it again turns sharply and becomes parallel to the axis when the second two-phase region is reached. After passing through this region the curves continue steeply upwards. The lines *aa*, *bb*, *cc* and *dd* joining the equivalent discontinuities on the series of curves, trace out the boundaries between the two-phase regions in the equilibrium diagram.

In the plot of $\log p$ against $1/T$ at constant concentration, the two-phase regions should be represented not by areas but by lines, since in these regions the pressure is independent of concentration. This can be seen to be true for the two-phase region represented by line *ff* in figure 3, but not for the two-phase region occurring at lower concentrations. This departure from ideal behaviour may be ascribed to superficial contamination of the titanium specimen by oxygen or nitrogen. Both these gases are known to have the effect of increasing the temperature at which the $\alpha \rightleftharpoons \beta$ transformation in titanium takes place. Thus contaminated parts of the specimen will transform at higher temperatures than the uncontaminated region. Since the effect of hydrogen itself is to reduce the transformation temperature, the purest material will transform at the lowest temperature. The line representing the two-phase region in the pure alloy should, therefore, pass through the discontinuities of lowest temperature. The slight departure from the horizontal in some of the pressure-concentration curves of figure 2 as they pass through the two-phase region is also an expression of this effect. A similar effect did not occur in the case of the high-concentration two-phase region *ff*, in which the alloy behaves ideally. This may be accounted for by the fact that the titanium specimens used in the determination of this part of the system were never heated above 600°C, whereas those used in establishing the lower concentration two-phase region were heated to temperatures as high as 950°C. The chances of gas contamination were, therefore, much smaller in the former case.

Experimental evidence in favour of the supposition that the alloy's departure from ideal behaviour in the low-concentration two-phase region is due to contamination by oxygen and nitrogen has been obtained from the study of a hydrogen-titanium alloy deliberately contaminated by allowing a small quantity of air to enter the apparatus. It was found that the low-temperature curves for this alloy fell on the curve *ee* of figure 3, but that the high-temperature discontinuity became less sharp and occurred at higher temperatures.

Since the pressure-temperature curves in the higher-temperature region of figure 3 are linear for the concentration involved in the $(\alpha + \beta)$ two-phase region, they may be produced to intersect the line through the lower-temperature discontinuities which represents the two-phase region for the pure material, as shown by the dotted lines in the diagram. The true pressure curves for the pure alloy can then be taken to follow round the curve *ee* until they leave it again at the points of intersection between the appropriate dotted lines and the curve *ee*. The effect of this correction on the pressure-concentration curves is indicated by the dotted lines in figure 2.

THE EQUILIBRIUM DIAGRAM

A limited region of the equilibrium diagram may be obtained directly from figure 2 by plotting temperature against concentration for each of the discontinuities

on the group of pressure-concentration curves. The part thus obtained is shown by the full lines in figure 4.

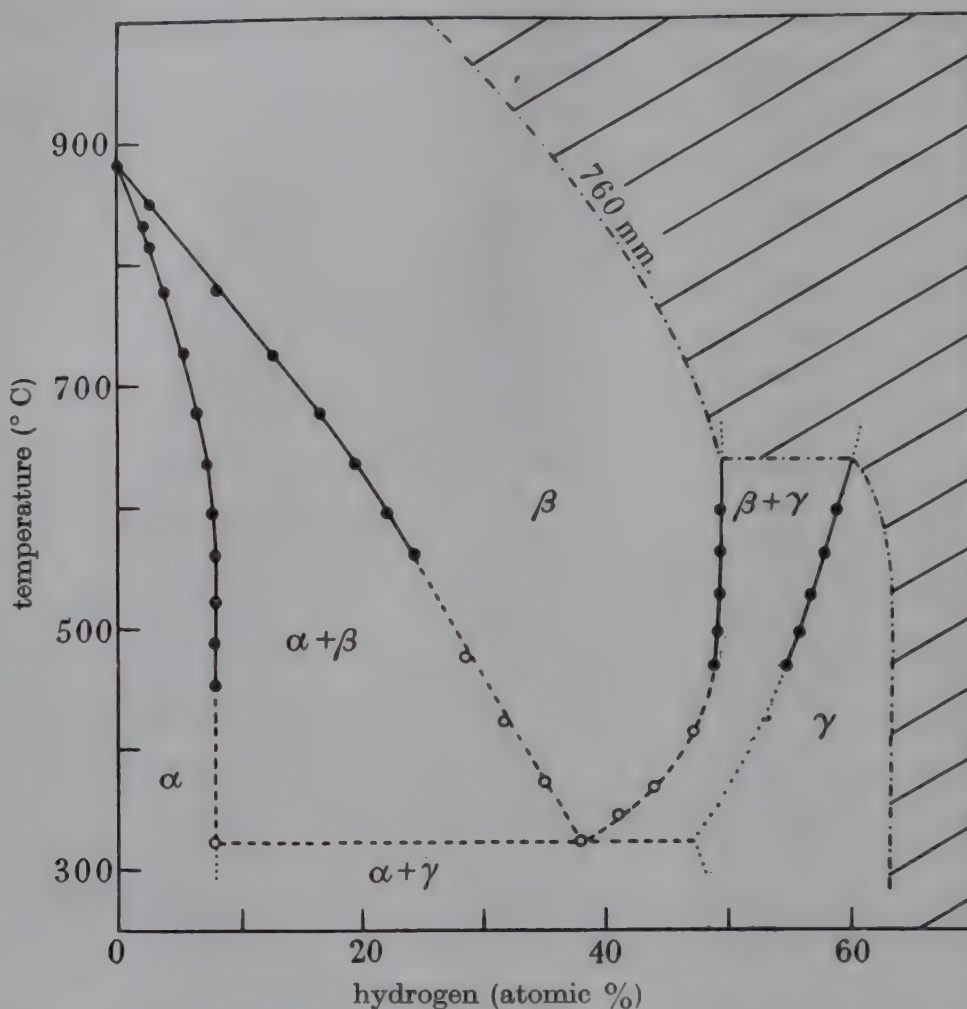


FIGURE 4. Hydrogen-titanium constitutional diagram for the condensed phases. In the figure, experimentally determined phase-boundaries are shown by a full line, while those obtained by extrapolation are indicated by a broken line. The dotted lines give the probable trends of the system. That part of the system for which the hydrogen pressure exceeds one atmosphere is denoted by a shaded area.

The boundary of the solid solution of hydrogen in the low-temperature hexagonal close-packed α -form of titanium is obtained from the line *aa* in figure 2. At higher hydrogen concentrations this single phase gives way to the two-phase region in which the α -solid solution co-exists with that based on the high-temperature body-centred cubic β -form of titanium. This is the two-phase region represented by the line *ee* in figure 3. The succeeding single-phase region is that of the β -solid solution. It will be observed that additions of hydrogen enable the body-centred cubic lattice of β -titanium to exist at temperatures much lower than those at which it exists in the pure metal. At still greater hydrogen concentrations a new phase, denoted by γ , appears. This has been found to have a face-centred cubic lattice. The boundary between the β and the $(\beta + \gamma)$ regions is represented by the line *cc* in figure 2 and the two-phase region itself by the line *ff* in figure 3. Alloys of the highest hydrogen content studied lie in the γ single-phase region, the beginning of which is given by the line *dd* in figure 2.

This experimental evidence is sufficient to indicate the general nature of the hydrogen-titanium system, which appears to be eutectoidal. By making several assumptions, however, it is possible to extend the diagram tentatively to temperatures lower than those for which direct evidence is available, thus obtaining a much more complete picture of the system.

The temperature-pressure curves for hydrogen-titanium alloys lying in the α - and β -regions are straight lines at the lowest temperatures, and, indeed, over most of the temperature range studied. It has also been found that, over the whole range of temperature covered by the experiments, the line *ff* representing the $(\beta + \gamma)$ region is straight, and a further linear relationship with temperature is shown by the line *ee* representing the $(\alpha + \beta)$ region, between 550 and 450° C. If it is assumed that the linearity of these relationships holds at temperatures below those which can be studied experimentally, points may be found, by producing the temperature-pressure lines in the region until they intersect a continuation of lines *ee* and *ff*, at which the alloys of the compositions represented by the various lines enter the two-phase regions $(\alpha + \beta)$ and $(\beta + \gamma)$. The point of intersection between the extrapolated lines *ee* and *ff* represents the unique point of contact between the $(\alpha + \beta)$ and $(\beta + \gamma)$ regions. It is, therefore, the eutectoid point.

Furthermore, the pressure-temperature curves in figure 3 are, in the α -region, exactly parallel to the linear part of the line *ee*. There must, therefore, exist an alloy in the α -region for which the pressure-temperature curve coincides with this linear part of the two-phase region curve. The composition of this alloy is found to be 7.9 atomic % hydrogen, and it represents the limiting concentration of the α -solid solution in the temperature range over which the line *ee* is known to be linear, i.e. 450 to 550° C. In this temperature range the $\alpha/(\alpha + \beta)$ boundary of the equilibrium diagram must, therefore, run parallel to the temperature axis. If the line *ee* is assumed to remain linear at temperatures below 450° C, the boundary must continue to be parallel to the temperature axis at a concentration of 7.9 atomic % until the eutectoid temperature is reached. This has been indicated in the tentative part of figure 4.

A further interesting point which emerges in the study of the equilibrium diagram is the form of the $\beta/(\beta + \gamma)$ phase boundary at higher temperatures. In the region of higher temperatures and pressures, the pressure-temperature lines for alloys in the β -condition change direction slightly, becoming more nearly parallel to line *ff* with increasing concentration (see figure 3). A concentration is eventually reached for which the pressure-temperature curve very nearly coincides with line *ff*, this concentration value being, therefore, the highest at which β can exist. As long as this pressure-temperature line remains parallel to the line representing the $(\beta + \gamma)$ region, the composition at which β changes to $(\beta + \gamma)$ will be independent of temperature—hence the almost vertical direction of the upper, high temperature, part of the $\beta/(\beta + \gamma)$ boundary in figure 4.

In spite of the seeming analogy between this effect and that relating to the low-temperature part of the $\alpha/(\alpha + \beta)$ boundary, the causes are different in the two cases. The rapid increase of pressure with concentration, shown by alloys in the β -condition having compositions approaching 50 atomic % hydrogen, indicates that it is

difficult for further amounts of hydrogen to enter the β -lattice at these compositions, and the alloys are almost completely saturated before the hydrogen pressure has risen to a value at which the γ -phase is stable. Additions of hydrogen to these alloys serve, therefore, to increase the external pressure rather than increase their composition.

Titanium has frequently been thought to form a hydride TiH_2 , supposedly a stoichiometric compound containing 66.6 atomic % of hydrogen. Analysis of the alleged compound usually indicates that its formula would be of the order of $\text{TiH}_{1.7}$. In figure 2 the pressure-concentration curves for the γ -phase rise steeply upwards and lie very close together. It would appear, therefore, that a limiting composition is reached beyond which increasing pressure has virtually no effect on the concentration. For pressures of the order of one atmosphere the limiting composition is 63.3 atomic % hydrogen corresponding to a formula $\text{TiH}_{1.73}$ —which is in good agreement with the general observation.

The equilibrium diagram was checked by micro-examination of alloys annealed at 750°C, slowly cooled to 300°C and held at this temperature for 24 hr. An alloy containing 5.1 atomic % hydrogen appeared as single phase α -solid solution, the 24.4, 36.6 and 43.7 atomic % alloys showed a eutectoid structure together with primary grains of either the α - or γ -phase, and the 61.7 atomic % alloy consisted of pure γ . The 36.6 atomic % alloy was almost entirely eutectoid, which compares well with the deduced value of 38 atomic % for the eutectoid point.

Lattice parameter measurements showed that though the presence of hydrogen had no measurable effect on the parameter of the α -titanium lattice, the parameter of the lattice of the γ -solution increased from 4.39₆ kX in the ($\alpha + \gamma$) two-phase region to 4.41₂ kX at 61.7 atomic %. It appears, therefore, that there is very little interaction between the hydrogen and titanium atoms in the α -solution, but that the interaction becomes appreciable in the γ -region.

MATHEMATICAL EXPRESSION OF THE PRESSURE-TEMPERATURE- CONCENTRATION RELATIONSHIPS

If it is assumed that the hydrogen molecules dissociate on going into solution in titanium and that the hydrogen-titanium system obeys the laws of dilute solutions, the hydrogen pressure p_{H_2} in equilibrium with a homogeneous condensed phase having a hydrogen concentration of c at a temperature T , will be given by

$$p_{\text{H}_2} = kc^2 \exp \frac{Q_0}{RT}, \quad (1)$$

where Q_0 is the heat of solution of molecular hydrogen in the metal, R is the gas constant and k is a constant, the value of which will depend on the structure of the condensed phase. Equation (1) accurately represents the behaviour of all α -solutions and of β -solutions containing less than 10 atomic % hydrogen. The values of Q_0 and k for these phases, when p_{H_2} is measured in mm. of mercury and c in atomic %, are given in table 1.

At hydrogen concentrations greater than 10 atomic %, equation (1) can no longer be applied to β -solutions. It is possible, however, to modify this equation and to

obtain an empirical expression which describes the pressure-temperature-concentration relationships in β -solutions at all temperatures and hydrogen concentrations for which the pressure does not exceed 150 mm. of mercury. This expression, which is accurate to within 1 %, is of the form

$$p_{\text{H}_2} = kc^2 \exp \lambda c^\mu \exp \frac{Q_0(1 + \eta c^\xi)}{RT}, \quad (2)$$

in which λ , μ , η and ξ are constants the values of which are presented in table 2. Equation (2) accounts both for the numerical increase in the heat of solution of hydrogen in β -phase alloys with increasing concentration shown by a progressive increase in slope of the $\log p_{\text{H}_2}/(1/T)$ curves of figure 3 as c increases, and for the rapid increase of p_{H_2} with c as c approaches 50 atomic %, shown in figure 2.

TABLE 1

	α -solution	β -solution (below 10 atomic %)
k	9.616×10^3	44.67×10^3
Q_0 (cal. per g.mol. of hydrogen)	-21.60×10^3	-27.83×10^3

TABLE 2

λ	1.331×10^{-4}
μ	2.638
η	9.433×10^{-6}
ξ	2.408

Since the pressure-temperature-concentration relationships for the γ -solutions cannot be related to the laws governing dilute solutions and the $\log p_{\text{H}_2}/(1/T)$ curves for this solution are not linear, it is not possible to deduce an analogous expression for the γ -solution.

THE FREE-ENERGY CURVES OF SOLUTIONS OF HYDROGEN IN α - AND β -TITANIUM

Since the activity of hydrogen at the pressures considered in these experiments may be taken to be proportional to the pressure, equations (1) and (2) may be used to calculate the partial free energy of hydrogen in α - and β -solutions. The partial free energy of titanium in each solution as a function of hydrogen concentration may then be deduced from the Gibbs-Duhem equation. The total free energy of both the α - and β - solutions at any temperature can, therefore, be given in terms of the free energies of pure α - and β -titanium at that temperature. It is then possible to obtain the difference between the free energies of α - and β -titanium as a function of temperature, over a limited temperature range, in the following way.

The free energy-concentration curves are constructed at a series of temperatures for both the α - and β -solutions of hydrogen in titanium. At each temperature, tangents are drawn to the α - and β - curves at points corresponding to the values of hydrogen concentration at which the α - and β -solutions co-exist. These values are obtained from the constitutional diagram. The tangents to the two curves are

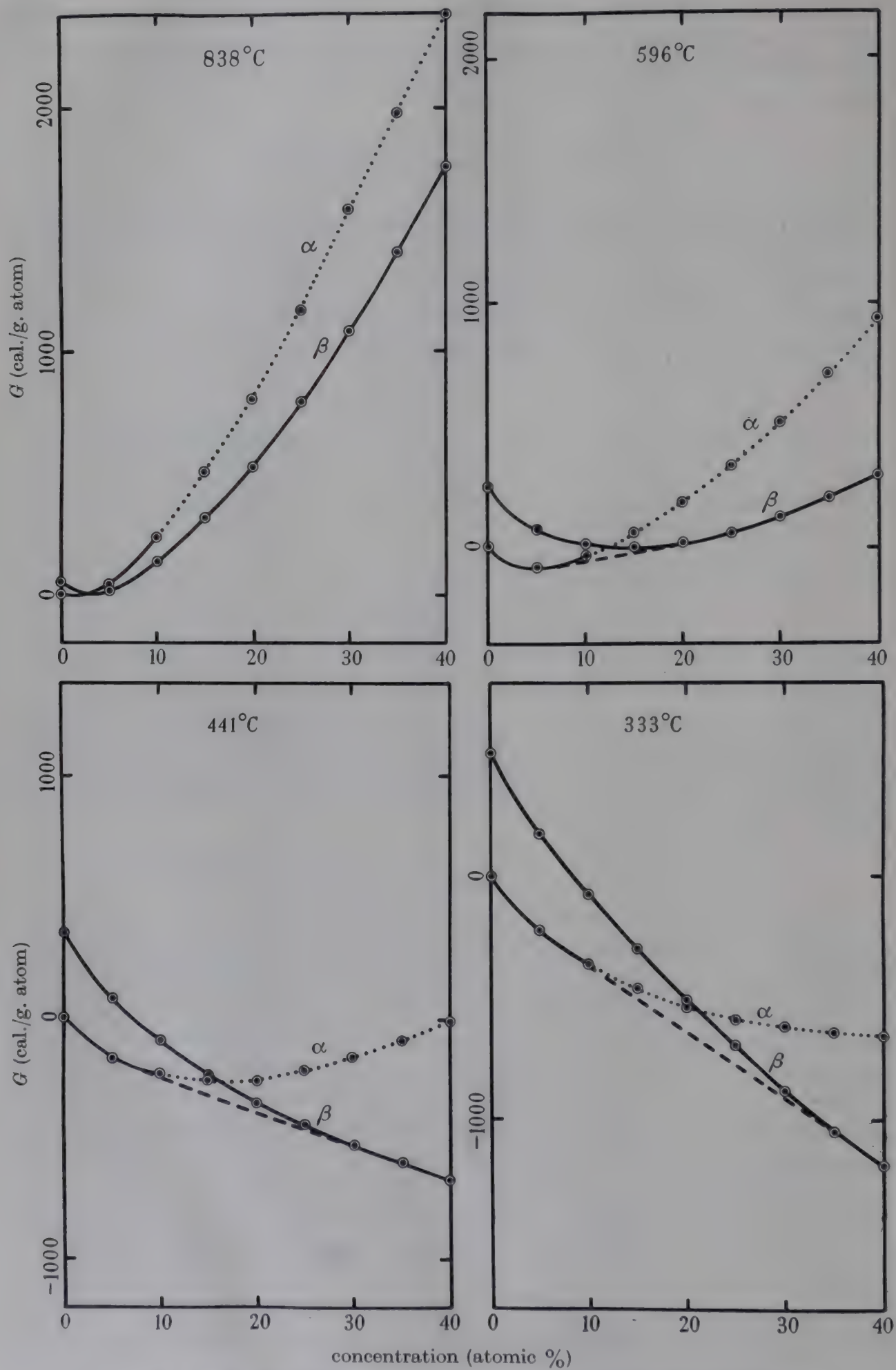


FIGURE 5. The free energy of hydrogen-titanium solutions as a function of concentration at a number of temperatures.

parallel, and may be made to coincide by a simple displacement of the free-energy curves in the direction of the free-energy axis. The extent of the displacement necessary to bring the tangents into coincidence gives the value of the difference between the free energies of α - and β -titanium at the temperature concerned. Figure 5 shows free-energy concentration curves for α - and β - solutions of hydrogen in titanium constructed for four different temperatures. The curves have been arranged so that the appropriate tangents coincide. The resulting relationship between temperature and free-energy difference between the α - and β -forms of pure titanium is given in figure 6. In the temperature region studied (882 to 333°C) the relationship appears to be linear, the rate of change of the free-energy difference with temperature being 0.86 cal./deg.

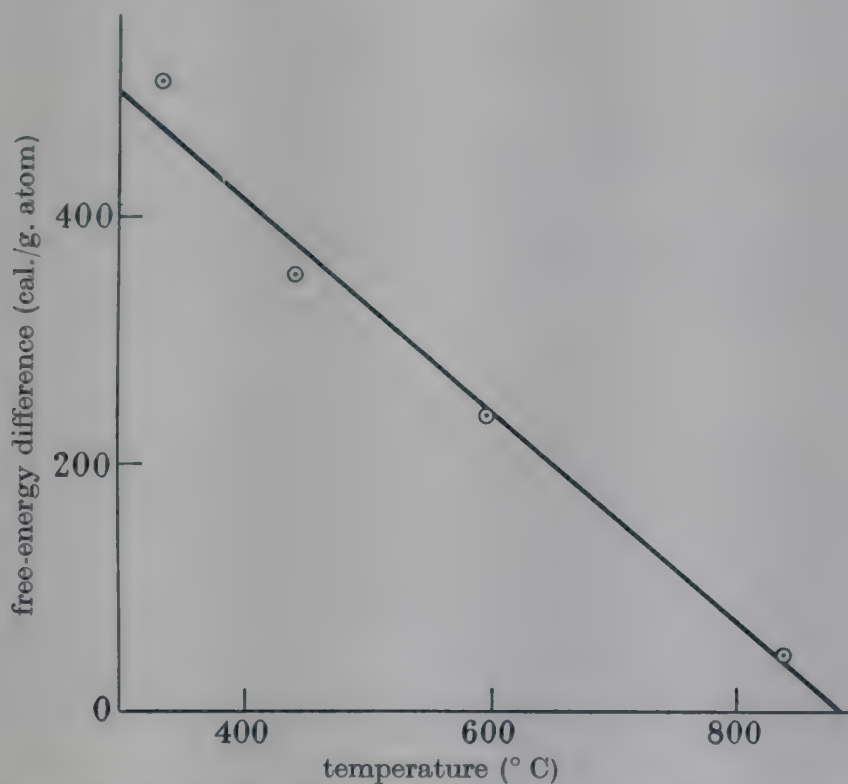


FIGURE 6. The free-energy difference between α - and β -titanium as a function of temperature.

THE TEMPERATURE AND LATENT HEAT OF THE $\alpha \rightleftharpoons \beta$ TRANSFORMATION IN PURE TITANIUM

From the known activities of titanium in α - and β -solutions for those hydrogen concentrations at which the two solutions are in equilibrium, the equilibrium constant, K , for the two forms of titanium may be determined as a function of temperature. This relationship is given in figure 7. The temperature at which $\ln K$ becomes zero corresponds to the transformation temperature of the pure metal and since the $\ln K/(1/T)$ curve becomes linear as $\ln K$ approaches zero it may be extrapolated to this value. The transformation temperature is found to be $882 \pm 0.5^\circ \text{C}$. The latent heat of transformation of titanium is proportional to the slope of the $\ln K/(1/T)$ curve at the transformation temperature and is found to be 678 cal./g. atom. The entropy change in the $\alpha \rightleftharpoons \beta$ transformation in titanium is, therefore, 0.587 cal./deg.

The values of the latent heat and entropy of the transformation in titanium may be compared with the values 710 cal./g. atom and 0.624 cal./deg. quoted by Smithells (1949) for the similar transformation which occurs in zirconium at 865° C. The values for titanium and zirconium are of the same order of magnitude, and in each case are much higher than those for the corresponding quantities for the transformations which occur in iron and manganese.

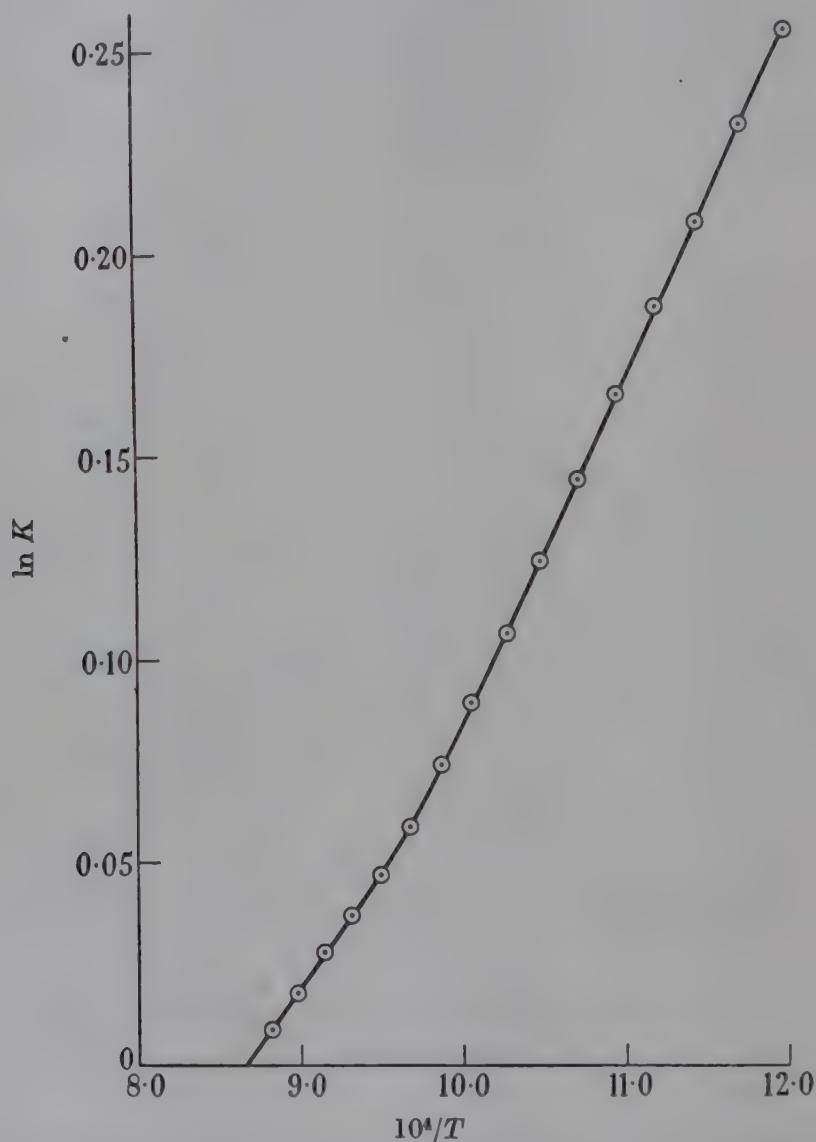


FIGURE 7. Equilibrium constant K for α - and β -titanium as a function of temperature.

CONCLUSION

It is hoped that the results obtained in this work which provide an example of the behaviour of a particular group of interstitial solutions, may contribute towards the better understanding of the mechanism of interstitial solution in transition metals and the statistical thermodynamic theory of gas-metal systems.

This work has been done as part of the research programme of the Physical Metallurgy Section of the Commonwealth Scientific and Industrial Research Organization, Melbourne, Australia.

REFERENCES

- Barrer, R. M. 1944 *Trans. Faraday Soc.* **40**, 374.
Fast, J. D. 1939 *Z. anorg. Chem.* **241**, 42.
Hägg, G. 1931 *Z. phys. Chem.* **11 B**, 433.
Kirschfeld, L. & Sieverts, A. 1930 *Ann. Phys. Chem.* **145 A**, 227.
Smithells, C. J. 1949 *Metals reference book*. London: Butterworth.
-

The scattering of gamma-rays in extended media II. Back-scattering of gamma-rays from a thick slab

BY J. CORNER AND R. H. A. LISTON

Ministry of Supply

(Communicated by W. G. Penney, F.R.S.—Received 22 June 1950)

The dose received by back-scattering at the front face of a thick layer of scattering material irradiated by a uniform beam of gamma-rays has been evaluated for beam directions between normal incidence and glancing incidence, for gamma-ray energies from 250 keV to 3 MeV. Results are given also for a half-slab, which is useful for building up re-entrant angles by superposition, and for a right-angled corner.

1. INTRODUCTION

In radiological work it often happens that there is scattering of radiation from neighbouring structures which it is inconvenient to remove, such as walls, floors and ceilings. While of course much of this scattered radiation can be removed by shields, it is usually desirable to keep the size and weight of shielding to a minimum. It is useful, therefore, to be able to estimate how much radiation is scattered back from walls. In the present paper we analyze two simple problems which frequently occur and which can be superposed to estimate the effect produced in more complicated layouts.

The theoretical methods available for scattering problems in extended media have been compared in part I by Cave, Corner & Liston (1950). In the present paper we use the simple method which they termed the 'modified first-order' approximation. This is sufficiently accurate for the present purpose, in which one needs only a little better than an order of magnitude. The 'modified first-order' method does in fact give an accuracy much better than this when, as in the present problems, the radiation which is to be evaluated travels only a small number of mean free paths in the scattering material.

In §2 we state the problems in more detail; in §3 we give the notation and in §§4, 5 and 6 the results for the whole plane, half plane and a right-angled corner respectively. An upper limit to the scattered dose is discussed in §7.

2. THE PROBLEMS STUDIED

We consider a wide parallel beam of monochromatic gamma-rays falling on a thick slab of the scattering material. We assume that the quantum energy is such

that photoelectric absorption and pair creation effects need not be taken into account. We have evaluated the scattered dose at energies from 250 keV to 3 MeV, for which range our assumption is justified for iron and all elements of lower atomic number. For substances of low atomic weight the energy range could be extended at the upper end, but here the scattered dose is so small that the extension is hardly worth while.

An advantage of considering only the Compton effect is that the results are independent of the scattering material.

There are of course other geometrical layouts in which one might want to assess back-scattering effects. It seems likely that the only important cases would be re-entrant angles. The simplest case, that of a right-angled concave corner, can be treated approximately by superposing a plane slab and half a plane slab at right angles. We have therefore computed also the back-scattered dose from half a slab, with which most of the interesting situations could be built up, approximately, by superposition.

3. NOTATION

The notation is indicated in figures 1 and 2. Let θ_0 be the angle of incidence of the primary quanta, measured from the normal to the slab face, and let θ be the emergent angle of a scattered quantum, measured from the normal. Let θ_1 be the deflexion of the quantum in a scattering, and ϕ the emergent azimuth angle, which is the angle between the planes ZOX and QOR in figure 2. We write also $\mu = \cos \theta$ and $\mu_1 = \cos \theta_1$. The value of $h\nu/mc^2$ for the incident quanta is written as γ , which is therefore approximately twice the quantum energy in MeV. Let $h\nu/mc^2$ for a once-scattered quantum be $E = \gamma/\{1 + \gamma(1 - \mu)\}$. We write σ_0 for the Klein-Nishina cross-section, per electron, for scattering of the primary beam, σ_1 for the Klein-Nishina cross-section for scattering of the once-scattered beam, and σ_a for the Klein-Nishina cross-section for energy absorption. Let N be the number of electrons per cm^3 of the scattering material and r_0 the classical radius of the electron $= e^2/mc^2 = 2.80 \times 10^{-13} \text{ cm}$. The

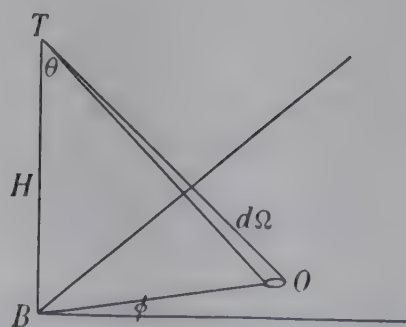


FIGURE 1

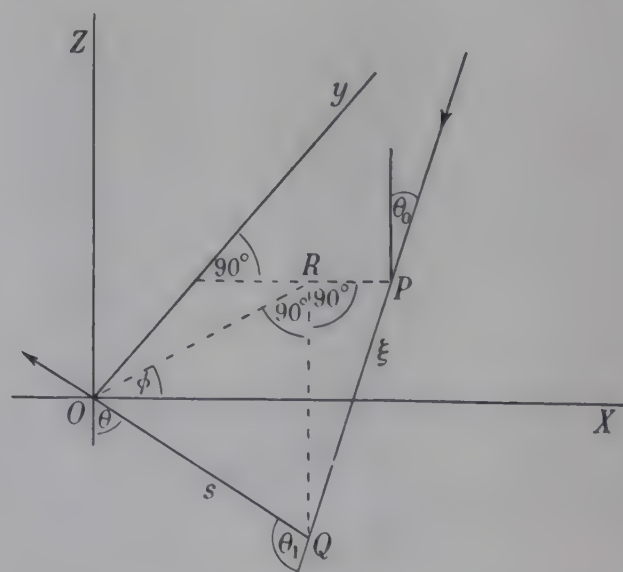


FIGURE 2

FIGURES 1 and 2. Notation for scattering problem.

probability for a quantum to be scattered in the direction μ , per unit solid angle, per unit initial intensity, per electron, we write as

$$g(\mu, \gamma) = \frac{r_0^2}{2} \left[\frac{1 + \mu^2}{[1 + \gamma(1 - \mu)]^2} + \frac{\gamma^2(1 - \mu)^2}{[1 + \gamma(1 - \mu)]^3} \right],$$

and it is convenient to define also the function

$$f(\mu, \gamma) = Eg(\mu, \gamma)/\gamma,$$

which is related to the energy scattered into direction μ .

4. DOSE FROM THE WHOLE PLANE

Let T in figure 1 be the point at which we measure the dose from rays scattered back from the block, whose dimensions parallel to its face are assumed to be very large. Let B be the point on the face directly under T .

An element of solid angle, $d\Omega = d\mu d\phi$, at T intercepts an area $H^2 d\Omega/\mu^3$ on the face. Let the number of quanta emitted from an area dA (around O) of the surface into a solid angle $d\Omega'$ round direction OT be $I(\mu, \phi) d\Omega' dA$ and let $h\nu/mc^2$ for these quanta be $E(\mu, \phi)$.

Take an element of volume $d\tau$ at T in the form of a cylinder with axis OT , cross-section $H^2 d\Omega'/\mu^2$ and length $\frac{\mu^2 d\tau}{H^2 d\Omega'}$. Of the energy emitted from the area $H^2 d\Omega/\mu^3$ at O the amount liberated in $d\tau$ at T is

$$N\sigma_a(E) E mc^2 I(\mu, \phi) \sec \theta d\Omega d\tau.$$

Let the gamma-rays incident on the slab have intensity one quantum per unit area. Their rate of energy liberation in $d\tau$ at T is then

$$N\sigma_a(\gamma) \gamma mc^2 d\tau$$

Thus back-scattering increases the dose by the relative amount

$$\int_0^1 \int_0^{2\pi} \frac{E}{\gamma} \frac{\sigma_a(E)}{\sigma_a(\gamma)} I(\mu, \phi) \sec \theta d\mu d\phi. \quad (1)$$

We now compute $I(\mu, \phi)$ for once-scattered quanta alone, and later we shall make an approximation to allow for multiple scattering.

Let scattering occur at distance s (in the range ds) from O and emerge in solid angle $d\Omega'$. The incident radiation comes in directions parallel to ZOX . The volume of the scattering element at Q is $d\tau' = s^2 ds d\Omega'$.

The incoming quanta must be scattered through a deflexion θ_1 in order to reach the surface at O , and to pass through dA they must be scattered into solid angle $\mu dA/s^2$. Let the scattering element at Q have cross-section dA' and length $d\tau'/dA'$ along PQ . This element receives $e^{-N\sigma_0 s} dA'$ quanta and scatters into solid angle $\mu dA/s^2$ a number

$$e^{-N\sigma_0 s} d\tau' Ng(\mu_1, \gamma) \mu dA/s^2,$$

so that the number reaching O is

$$e^{-N\sigma_1 s - N\sigma_0 s} Ng(\mu_1, \gamma) \mu d\Omega' ds dA.$$

Now the number of quanta emitted from dA in the surface into solid angle $d\Omega'$ around OT has been written as $I(\mu, \phi) d\Omega' dA$. Hence

$$I(\mu, \phi) = N\mu g(\mu, \gamma) \int_0^\infty e^{-N\sigma_1 s - N\sigma_0 \xi} ds, \quad (2)$$

and substituting this in (1) we have the relative increase in dose,

$$N \int_0^1 \int_0^{2\pi} \frac{E}{\gamma} \frac{\sigma_a(E)}{\sigma_a(\gamma)} g(\mu_1, \gamma) d\mu d\phi \int_0^\infty e^{-N\sigma_0 \xi - N\sigma_1 s} ds. \quad (3)$$

Since

$$s \cos \theta = RQ = \xi \cos \theta_0,$$

we have

$$N \int_0^\infty e^{-N\sigma_0 \xi - N\sigma_1 s} ds = 1/(\sigma_1 + \sigma_0 \mu / \cos \theta_0),$$

and the relative increase of dose can therefore be written as

$$\frac{1}{\sigma_a(\gamma)} \int_0^1 \int_0^{2\pi} \frac{\sigma_a(E) f(\mu_1, \gamma)}{\sigma_1 + \sigma_0 \mu / \cos \theta_0} d\mu d\phi. \quad (4)$$

In this expression

$$\cos \theta_1 = \mu_1 = -\mu \cos \theta_0 + \sin \theta_0 \sin \theta \cos \phi. \quad (5)$$

We now make the approximations of the 'modified first-order' method (Cave *et al.* 1950):

$$\sigma_1 = \sigma_0, \quad \sigma_a(E) = 1.243r_0^2 \text{ for } \gamma \geq 1, \quad \sigma_a(E) = \sigma_a(\gamma) \text{ for } \gamma \leq 1.$$

Thus the relative increase in dose becomes

$$\left\{ \frac{\sigma_a(E)}{\sigma_a(\gamma)} \right\} \frac{1}{\sigma_0} \int_0^1 \int_0^\pi \frac{2f(\mu_1, \gamma)}{1 + \mu \sec \theta_0} d\mu d\phi. \quad (6)$$

The integral has been computed for $\gamma = 0.5, 1, 4$ and 6 and the results are shown in figure 3.

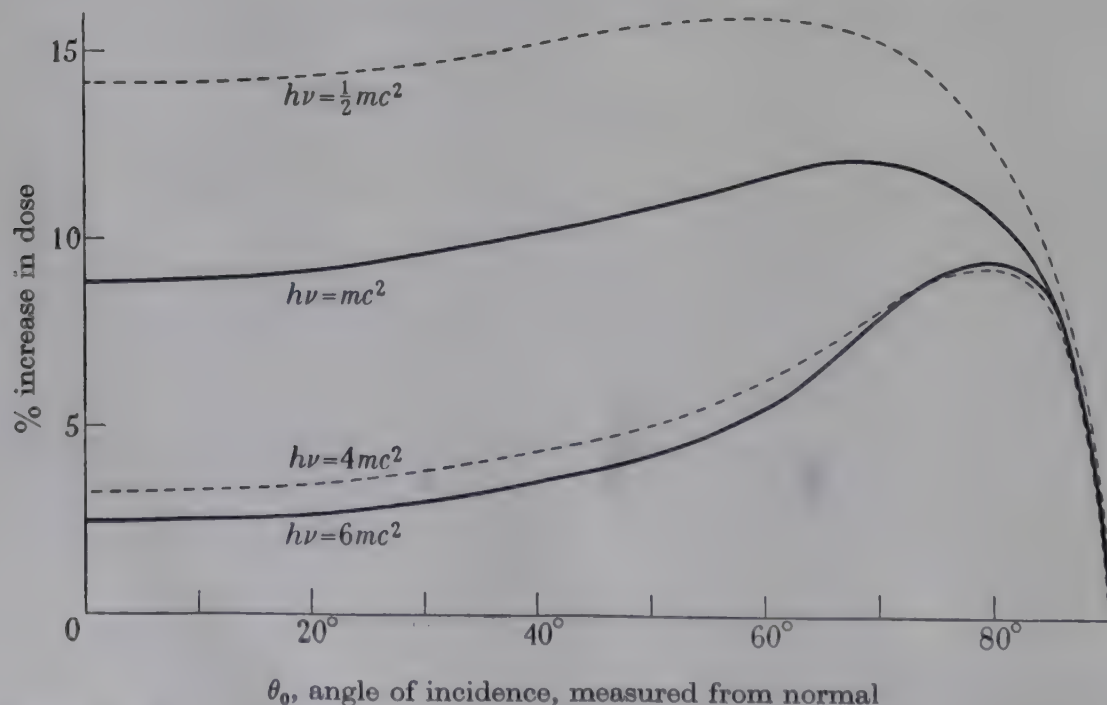


FIGURE 3. Dose by scattering back to upper medium from a thick slab of large area.

It will be noticed that the relative increase in dose for $\gamma = 6$ is slightly greater than for $\gamma = 4$ in the range of incident angles from 75° to 90° . This would not be the case in practice, where obviously the dose should be slightly less. The discrepancy is due to the approximation of replacing $\sigma_a(E)$ by $1.243r_0^2$ whatever the value of γ .

5. THE HALF-PLANE

We now consider the case where the scattering medium is cut off at the vertical plane through the point at which the dose is measured. The only change in (6) is that ϕ is now restricted to the range $-\frac{1}{2}\pi$ to $\frac{1}{2}\pi$.

The expression (6) has been integrated over these limits and the results for the half plane are plotted in figure 4.

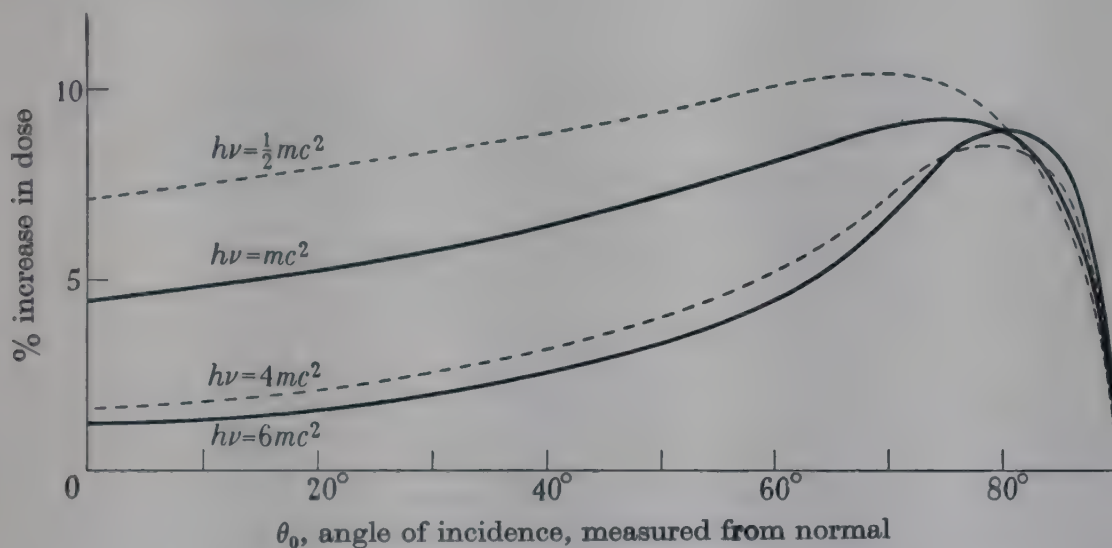


FIGURE 4. Dose by scattering back to upper medium from a thick slab, semi-infinite in extent.

6. SCATTERED RADIATION IN A CORNER

If we can tolerate an overestimate for scattering paths such as ABT in figure 5 (the probability of such a path is small even at $\gamma = 0.5$), we may add the result for a whole plane for incidence θ_0 with the result for a half plane with incidence $(\frac{1}{2}\pi - \theta_0)$ to find

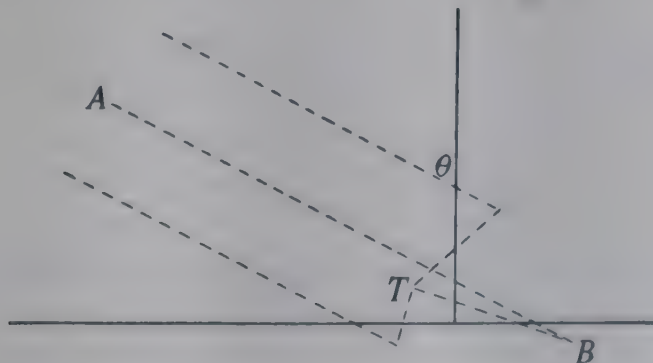


FIGURE 5. Paths for scattering of gamma-ray quanta towards a counter in a right-angle.

the relative increase in dose at a point lying inside a right-angle formed by two planes.

When we perform this addition the curve obtained for dose against θ_0 is slightly asymmetrical. Physical consideration will show that the true curve should be symmetrical about $\theta_0 = \frac{1}{4}\pi$. We have accordingly averaged the values for θ_0 and $(\frac{1}{2}\pi - \theta_0)$ to obtain a symmetrical curve. This procedure will introduce a small error at all values of θ_0 , but this is unimportant in view of the errors from (i) multiple scattering not being taken into account exactly, and (ii) overestimate of scattering from certain types of path. The asymmetries which were removed were in fact of order only 2 % of the primary dose.

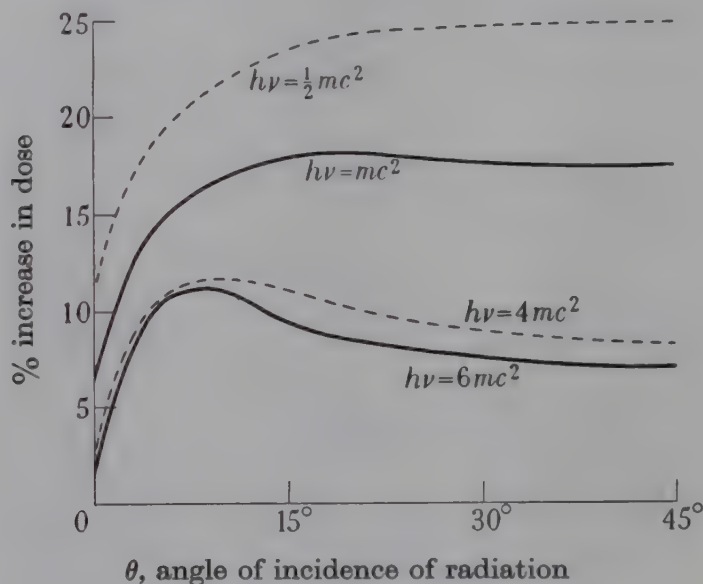


FIGURE 6. Scattered dose received in a right-angled concave corner.

It is interesting to note that for a given value of γ the increase in dose varies by only a few per cent over practically the whole range of incident angles.

7. THEORETICAL UPPER LIMIT TO THE INCREASE IN DOSE

It is of some interest to know the maximum increase in dose to which back-scattering can give rise. We may find this maximum by considering the case where $\gamma = 0$, that is, the case of pure Thomson scattering.

Under these conditions the probability function $2f(\mu, \gamma)/r_0^2$ is replaced by $(1 + \mu_1^2)$ and σ_0/r_0^2 takes the limiting value $8\pi/3$. The expression (6) now becomes

$$\frac{3}{8\pi} \int_0^1 \int_0^\pi \frac{1 + \mu_1^2}{1 + \mu \sec \theta_0} d\mu d\phi.$$

Consider the case when $\theta_0 = 0$ (perpendicular incidence). Then from (4) $\mu_1 = -\mu$ and we have for the relative increase of dose (whole plane)

$$\frac{3}{8\pi} \int_0^1 \int_0^\pi \frac{1 + \mu^2}{1 + \mu} d\mu d\phi = \frac{3}{8} \int_0^1 \frac{1 + \mu^2}{1 + \mu} d\mu = \frac{3}{8} (2 \log_e 2 - \frac{1}{2}) = 0.3324.$$

The limiting value could be worked out for other values of θ_0 , but this is hardly necessary, since it is clear from the behaviour of the results for various values of γ that the curve will be practically flat and the maximum theoretical increase will not exceed about one-third.

We are indebted to Dr W. G. Penney, F.R.S. for his encouragement of this work, and to the Chief Scientist, Ministry of Supply, for permission to publish.

REFERENCE

Cave, L., Corner J. & Liston, R. H. A. 1950 *Proc. Roy. Soc. A*, **204**, 223 (part I).

The scattering of gamma-rays in extended media III. Problems with spherical symmetry

BY J. CORNER, F. A. G. DAY AND R. E. WEIR
Ministry of Supply

(Communicated by W. G. Penney, F.R.S.—Received 22 June 1950)

Some typical problems of the propagation of gamma-radiation in an extended scattering medium have been solved for pure Compton scattering, which is a good approximation for most materials at gamma-ray energies up to a few MeV. The effects to be expected from air gaps round the source, large source dimensions, and plane absorbing materials are demonstrated.

1. INTRODUCTION

In an earlier paper, part I (Cave, Corner & Liston, 1950), mathematical methods for handling repeated scattering of gamma-rays in extended media were tested on a particularly simple example, and it was shown that there were two rapid methods which gave surprisingly good results. In the present paper we apply these methods to some problems with spherical symmetry, chosen as simple idealizations of situations which occur in radiography.

For substances of medium and low atomic weight, the process which is most important is Compton scattering. This is true, at any rate, from quantum energies of order 200 keV up to the highest energies, of a few MeV, available from natural sources. In the present paper, therefore, we neglect pair-creation and photoelectric absorption. This approximation is convenient because a scattering medium is then described by a single parameter, the number of electrons, free and bound, per unit volume.

In §2 we show some typical results for a point source in an infinite homogeneous medium, with an indication, in §3, of how little the dose is altered if the source is distributed throughout a sphere, provided this has the same scattering properties as

the rest of the medium. In later sections we show how the field is affected by non-scattering media, in the form of (§ 4) a semi-infinite medium with a plane face not too near the source, and, in § 5, an empty sphere round the point source. Mathematical details are given in two appendices.

2. THEORETICAL RESULTS FOR A POINT SOURCE IN A HOMOGENEOUS MEDIUM

We can classify the radiation arriving at a point into that which has not been scattered, that which has been scattered once, and so on. It is easy to compute the dose from unscattered and once-scattered quanta. The two best of the reasonably easy approximate methods tested by Cave *et al.* (1950) both take the integral for the once-scattered radiation and modify it to allow for higher-order scatterings. We have applied both these methods to the problem of the point source in a homogeneous medium. The theory is given in appendices A and B. The computation involves nothing worse than numerical double integration.

Let r be distance from the source, and let D be the distance measured in mean free paths for the initial quanta, which have $h\nu/mc^2 = \gamma$. Thus $D = N\sigma_0 r$, where N is the number of electrons, both free and bound, per cm^3 , and σ_0 is the total Compton scattering coefficient (per electron) at $h\nu/mc^2 = \gamma$. The dose at distance r can be represented as $r^{-2}e^{-\alpha D}$ multiplied by a parameter depending only on the number of quanta emitted by the source and on the value of γ . This parameter is of no interest in the present report, and we omit it. The other function, α , is a number depending on γ and D alone and is of order 0.7. The ‘apparent mean free path’ is $1/\alpha$ times the true mean free path $1/N\sigma_0$.

The two theoretical methods are compared at $\gamma = 4$, that is, at $h\nu \simeq 2 \text{ MeV}$, in table 1, which includes also the results for perpendicular incidence on a plane slab; here the dose, relative to that at the front of the slab, is written as $e^{-\alpha D}$. A fairly close correspondence between the plane and spherical cases can be anticipated, and the comparison gives a check on the results. Undoubtedly the ‘modified upper limit’ method of appendix B is the better of the two. The ‘modified first-order’ (appendix A) is useful chiefly because it is the easier to apply in more complicated geometrical conditions.

It is clear from table 1 that the difference between ‘plane’ and ‘spherical’ calculations is small, and indeed smaller than the real error in these simple methods, and the extra labour involved in the ‘spherical’ computation suggests that ‘plane’ solutions are adequate in most cases.

TABLE 1. VARIATION OF DOSE WITH DISTANCE IN PLANE AND SPHERICAL PROBLEMS

$\gamma = 4$; dose $\propto e^{-\alpha D}$ for perpendicular incidence on a plane slab: dose $\propto r^{-2} e^{-\alpha D}$ for spherical symmetry; D = distance in mean free paths = $N\sigma_0 r$. Values of α are tabulated.

D	‘modified upper limit’		‘modified first-order’ method	
	plane	spherical	plane	spherical
3	0.667	0.673	0.702	0.693
5	0.712	0.719	0.774	0.758
8	0.747	0.751	0.830	0.813

The value of α varies slowly with γ in the region of interest in radiography, as is shown by table 2.

TABLE 2. VARIATION OF α WITH $h\nu/mc^2$, FROM 'MODIFIED UPPER LIMIT' THEORY

D	3	5	8
α at $\gamma = 4$	0.673	0.719	0.751
α at $\gamma = 6$	0.679	0.726	0.763

3. RESULTS FOR A BALL SOURCE

It is easy to allow for the gamma-ray sources being distributed through a small sphere with the same scattering properties as the rest of the medium. The formulae are given in appendix A, § A 2, and appendix B, § B 2. Computations were carried out by the modified first-order method for $\gamma = 4$, $D = 3$, with a spherical source of radius one half a mean free path, which is as large as is likely to be used. It was found that α was altered by less than 0.001. At larger distances the effect is obviously even smaller. Thus the effect is negligible except at distances of order one mean free path, a situation which is not of much interest in radiography.

4. A POINT SOURCE NEAR A PLANE INTERFACE WITH A LARGE CHANGE OF MEAN FREE PATH

It sometimes happens that we have two media separated by a plane interface, at which we require to know the dose when there is a gamma-ray source inside one of the media. In practice the commonest case has air as one of the media, with a considerable difference in the mean free paths in the two substances. Such a problem can be treated very conveniently by finding the effect when the source-free medium is a vacuum. If this medium has the relatively long mean free path the approximation is evidently a good one. If this medium has the short mean free path we must allow also for back-scattering, which will come only from points close to the point of measurement. This correction can be obtained from the results of the preceding paper (Corner & Liston 1950), on back-scattering of a parallel beam from a plane slab. It is only necessary to choose an 'average' direction of incidence of the gamma-rays reaching the interface near the point of observation, and this can be done from the results for a point source in a homogeneous medium; since the back-scatter is not sensitive to the angle of incidence except near glancing incidence, the choice of the 'average' direction is not important.

TABLE 3. TYPICAL CHANGE IN DOSE DUE TO VACUUM BEYOND A CERTAIN PLANE

Point source distant one mean free path from interface. D is distance to source, in mean free paths. $\gamma = 4$.

D	3	8
dose with vacuum/dose for infinite medium	0.956	0.849

When the dose is required at a point which lies on a plane interface with a vacuum, there are some simple changes in the 'modified first-order' result (appendix A, § A 3). Typical results are shown in table 3, where $\gamma = 4$ and the distance of the point source from the interface, H mean free paths, is $H = 1$. Smaller distances are not of much

interest. It is seen that the effects are only a few per cent for distances of a few mean free paths. At very large distances, of course, the dose must be roughly halved, but this ratio would evidently not be approached under normal conditions.

5. A GAP ROUND THE SOURCE

In practice there is often a gap, of order one mean free path wide, around the source, and it is interesting to see what effect this has on the radiation at points in the otherwise homogeneous medium.

When a spherical hole is made in a scattering medium the unscattered radiation arriving at a point outside the hole is increased by a factor e^Y , where Y is the radius of the hole in mean free paths. The scattered radiation also travels a shorter distance in the scattering medium, but this is partly offset by the loss of radiation which was formerly scattered by the material in the hole. The net result differs in the two approximate solutions we have used. From the ‘modified first-order’ approximation it can be shown that for Y of order unity and D of order a few units the scattered dose is nearly proportional to $e^Y(1 - Y/D)$, in its dependence on Y at a given D . If we write the factor for the scattered dose as $e^Y(1 - \theta Y/D)$ in the ‘modified upper limit’ solution, then θ is found to vary with Y by only a few per cent for Y up to $\frac{1}{2}D$, and to change by not much more than 10 % when γ is altered from, say, 4 to 6. θ does, however, increase with D , typical values being $\theta = 1.1$ at $D = 3$ and $\theta = 1.5$ at $D = 8$. Since Y/D does not usually exceed one-third, the differences between the two methods are not important. Table 4 shows how closely the factor $e^Y(1 - Y/D)$ reproduces typical values computed from the ‘modified first-order’ approximation (§ A 4 of appendix A) and gives for comparison some results from the ‘modified upper limit’ (appendix B, § B 3), which is probably more accurate. The spectral distribution of scattered radiation is little altered by the spherical hole. This is true in both approximations.

TABLE 4. INCREASE IN SCATTERED DOSE DUE TO PRESENCE OF A HOLE OF RADIUS Y MEAN FREE PATHS

γ	D	Y	scattered dose/scattered dose for $Y = 0$		$e(1 - Y/D)$
			‘modified upper limit’	modified first- order’	
4	3	1	1.677	1.836	1.812
4	8	1	2.129	2.340	2.378

We are indebted to Dr W. G. Penney, F.R.S., for his encouragement of this work, and to the Chief Scientist, Ministry of Supply, for permission to publish.

APPENDIX A. MATHEMATICAL SOLUTION OF PROBLEMS WITH SPHERICAL SYMMETRY BY THE ‘MODIFIED FIRST-ORDER’ APPROXIMATION

A 1. Point source, homogeneous atmosphere

Figure 1 illustrates the notation. A point source at O emits quanta isotropically into an infinite medium containing N electrons, free or bound, per cm.³. The value

of $h\nu/mc^2$ for the initial quanta is γ ; here m is the electron mass and c is the velocity of light, so that $mc^2 = 0.511$ MeV. Let the Compton scattering cross-section per electron for the initial quanta be σ_0 . Let $\sigma_a(E)$ be the Compton cross-section per electron, for absorption of energy from quanta with $h\nu/mc^2 = E$. We neglect all processes except Compton collisions.

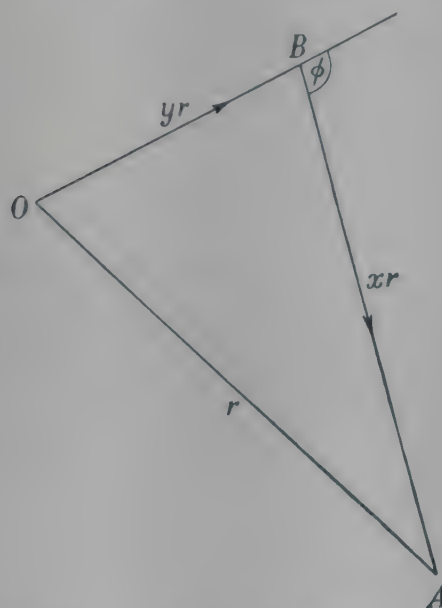


FIGURE 1. Notation for scattering problem.

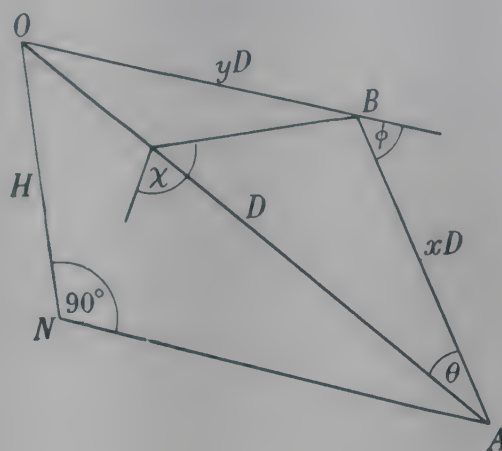


FIGURE 2. Notation for scattering near a plane interface.

In a radiation field in which the intensity depends on direction and $E = h\nu/mc^2$, the energy absorbed per electron in a small element of a substance of low atomic weight is

$$mc^2 \iint E \sigma_a(E) I(\mathbf{n}, E) d\Omega dE, \quad (1)$$

where $\mathbf{n}$ is a unit vector, I is the intensity in direction $\mathbf{n}$ and at $h\nu/mc^2 = E$, and the integrations are taken over all directions and the whole spectrum of E . The intensity I is defined by the number of quanta crossing a small area with directions of motion in a small solid angle, with $h\nu/mc^2$ lying in a small band dE , and has the dimensions of a number of quanta per unit area, unit solid angle, and unit spread in E .

It can be shown that for whole-body irradiation the general physiological effects can be correlated with (1).

The chief interest lies in the decay of (1) with distance from the source. It is convenient, therefore, to divide throughout by $mc^2 \gamma \sigma_a(\gamma)$ and to state everything 'per quantum emitted'. The 'standard dose' so defined, namely

$$\text{s.d.} = \iint (E/\gamma) \{ \sigma_a(E) / \sigma_a(\gamma) \} I d\Omega dE \quad (2)$$

could easily be converted to Röntgens if need be. If there were no attenuation by scattering, the 'standard dose' at distance r would be $1/4\pi r^2$. If no scattered quanta contributed to the dose, the result would be $e^{-D}/4\pi r^2$, where $D = N\sigma_0 r$. In practice the truth will lie somewhere between these extremes, and we write the standard dose as $e^{-\alpha D}/4\pi r^2$, where α is a number between 0 and 1, depending on γ and D .

We proceed to evaluate the standard dose due to quanta scattered once before arriving at the point A where the measurements are taken. Let $OA = r$. The quanta travel a distance yr from O to B where they are scattered through an angle ϕ and travel a distance xr from B to A . Along BA the quanta have a value of $h\nu/mc^2 = E_1$, where

$$E_1 = \gamma / \{1 + \gamma(1 - \cos \phi)\}, \quad (3)$$

at which energy the Compton scattering cross-section per electron is σ_1 , say.

Directions at A can be specified by the angle θ made with AO . Let $\mu = \cos \theta$. Then

$$y^2 = 1 + x^2 - 2x\mu, \quad (4)$$

and

$$\cos \phi = (\mu - x)/y. \quad (5)$$

Take a small solid angle $d\Omega$ around AB and consider quanta which are scattered in a volume $d\tau = r^3 x^2 dx d\Omega$ at B . The probability that a quanta will travel from O to B without a scattering, will be scattered in $d\tau$ into directions which cross a small area dA at D , and will then travel the distance BA without further scattering, is

$$\frac{e^{-N\sigma_0 r y}}{4\pi(yr)^2} N d\tau g(\cos \phi, \gamma) \frac{dA}{(xr)^2} e^{-N\sigma_1 x r} = \frac{NdA d\Omega}{4\pi r} y^{-2} e^{-Dy - D\sigma_1 x/\sigma_0} g(\cos \phi, \gamma) dx, \quad (6)$$

where $g(\cos \phi, \gamma)$ is the Compton coefficient for scattering of quanta into a given direction

$$g(\cos \phi, \gamma) = \frac{r_0^2}{2} \left[\frac{1 + \cos^2 \phi}{\{1 + \gamma(1 - \cos \phi)\}^2} + \frac{\gamma^2(1 - \cos \phi)^2}{\{1 + \gamma(1 - \cos \phi)\}^3} \right], \quad (7)$$

and $r_0 = e^2/m$ is the classical 'radius of the electron'. Integrating (6) over all positive values of x , we obtain $IdA d\Omega$, normalized to unit quantum emitted. Using this intensity in the 'standard dose' and noting that E takes only the value E_1 given by (3), we find for the contribution from once-scattered quanta

$$\frac{N}{4\pi r} \int_{\mu=-1}^1 \int_{x=0}^{\infty} y^{-2} e^{-Dy - D\sigma_1 x/\sigma_0} \frac{E_1}{\gamma} \frac{\sigma_a(E_1)}{\sigma_a(\gamma)} g(\cos \phi, \gamma) dx 2\pi d\mu,$$

and writing

$$f(\cos \phi, \gamma) = (E_1/\gamma) g(\cos \phi, \gamma), \quad (8)$$

so that $f(\cos \phi, \gamma)$ is the Compton cross-section for the scattering of *energy* through a deflexion ϕ , we obtain for the standard dose from unscattered and once-scattered radiation

$$\text{S.D.} = \frac{e^{-D}}{4\pi r^2} \left[1 + \frac{2\pi D e^D}{\sigma_0 \sigma_a(\gamma)} \int_{\mu=-1}^1 \int_{x=0}^{\infty} y^{-2} e^{-Dy - D\sigma_1 x/\sigma_0} \sigma_a(E_1) f(\cos \phi, \gamma) dx d\mu \right]. \quad (9)$$

Some correction must now be made for radiation which arrives at A after more than one scattering. In the modified first-order method this correction is made, rather crudely but apparently quite effectively, by writing $\sigma_1 = \sigma_0$ and by replacing $\sigma_a(E_1)$ in (9) by the maximum value $\sigma_a(1) = 1.2430r_0^2$ if $\gamma \geq 1$ and by $\sigma_a(\gamma)$ if $\gamma < 1$. The standard dose becomes

$$\frac{e^{-D}}{4\pi r^2} \left[1 + \frac{2\pi D e^D \times 1.243r_0^2}{\sigma_0 \sigma_a(\gamma)} \int_{\mu=-1}^1 \int_{x=0}^{\infty} y^{-2} e^{-D(x+y)} f(\cos \phi, \gamma) dx d\mu \right], \quad (10)$$

if $\gamma \geq 1$, which is the more important case.

It is useful to be able to see how the scattered radiation is distributed from $h\nu/mc^2 = \gamma$ to $\gamma/(1+2\gamma)$. This suggests that the double integration should be performed first with respect to x or μ , then with respect to E_1 . This is convenient also in reducing the number of times the awkward function $f(\cos \phi, \gamma)$ has to be computed.

The standard dose from quanta with E_1 lying in a range dE_1 , relative to the dose from primary quanta, is, by (3) and (10),

$$\frac{2\pi D e^D 1.243 r_0^2 f(\cos \phi, \gamma)}{\sigma_0 \sigma_a(\gamma)} \int_{x=0}^{\infty} y^{-2} e^{-D(x+y)} dx d\mu,$$

where $d\mu$ is the spread of values of μ which for a given value of x have $\cos \phi$ lying in a range $E_1^{-2} dE_1$. From (4) and (5),

$$d\mu = y^3 E_1^{-2} dE_1 / |1 - x\mu|,$$

so that the total standard dose becomes

$$\text{s.d.} = \frac{e^{-D}}{4\pi r^2} \left[1 + \frac{2\pi D e^D 1.243 r_0^2}{\sigma_0 \sigma_a(\gamma)} \int_{\gamma/(1+2\gamma)}^{\gamma} E_1^{-2} f(\cos \phi, \gamma) dE_1 \int_0^{\infty} y e^{-D(x+y)} dx / |1 - x\mu| \right]. \quad (11)$$

For ϕ between 0 and 90° , corresponding to E_1 between $\gamma/(1+\gamma)$ and γ , there are real solutions of (4) and (5) from $x = 0$ to 1 , and the only difficulty which appears to arise in the computation is that y and $1 - x\mu$ tend to zero as x approaches unity. However, their ratio approaches the finite limit $(1 + 1/\gamma - 1/E_1)^{-1}$.

For ϕ between 90° and 180° , so that E_1 lies between $\gamma/(1+2\gamma)$ and $\gamma(1+\gamma)$, x starts at O with $\mu = \cos \phi$, and as μ increases towards $+1$, x first increases to a value greater than 1 , then reverses and returns to $x = 1$. In the neighbourhood of $x = \mu = 1$, y and $1 - x\mu$ again become zero and their ratio tends to $|1 + 1/\gamma - 1/E_1|^{-1}$. The only other difficulty occurs at the point where x reverses, that is, where $dx/d\mu = 0$. Here $1 - x\mu = 0$ and the integrand is infinite. If we write $h(x)$ as an abbreviation for the well-behaved function $y e^{-D(x+y)}$, we find that

$$\int_X^{x_0} h(x) dx / |1 - x\mu| = x_0 h(x_0) [2(1 - X/x_0)]^{\frac{1}{2}} |1 + 1/\gamma - 1/E_1|^{-1}, \quad (12)$$

if x_0 is the maximum value attained by x , actually

$$x_0 = [1 - (1 + 1/\gamma - 1/E_1)^2]^{-\frac{1}{2}}, \quad (13)$$

and X is any value of x close to x_0 . In practice $h(x)$ can vary fairly rapidly in the range from X to x_0 . For a linear variation, which is usually acceptable, $h(x_0)$ in (12) should be replaced by $\{2h(x_0) + h(X)\}/3$.

For quantum energies of a few MeV it was found that it was necessary to use at most one value of E_1 giving negative $1 + 1/\gamma - 1/E_1$ and so leading to such difficulties. Here we are not counting the lower limit $E_1 = \gamma/(1+2\gamma)$, for which the integration can be carried out analytically. The region of integration falls into two parts: in the first, $\mu = -1$, $y = 1 + x$, x goes from 0 to infinity, and

$$\int_0^{\infty} y e^{-D(x+y)} dx / |1 - x\mu| = e^{-D/2D};$$

in the second zone, $\mu = 1$, $y = x - 1$, x goes from 1 to infinity, and the integral again gives $e^{-D}/2D$.

At the upper limit $E_1 = \gamma$ the integration can also be carried out analytically.

A 2. Spherical source in homogeneous medium

A sphere is surrounded by an infinite medium with the electron density N the same in the sphere as outside. The sphere contains a uniform distribution of gamma-ray sources, and we require the dose outside the sphere.

From the solution for a point source it is easy to write down the solution for the spherical source. The unscattered radiation can be integrated explicitly in terms of exponential integrals and hyperbolic functions, while the scattered radiation appears as a double integral somewhat more complicated than (10). This could be computed much as before. It appears, however, to be quicker to compute (11) as a function of D , weight the result by the distribution of the sources over D , and finally integrate numerically.

Let zr be the radius of the source and, for short, let us write (11) as $s.D.(r, D)$. Then the standard dose from the spherical source, per quantum emitted, is

$$\frac{3}{4z} \int_{1-z}^{1+z} s.D.(rs, Ds) [1 - (1-s)^2 z^{-2}] s ds. \quad (14)$$

A 3. Vacuum beyond a plane interface, point source

The notation is shown in figure 2, in which the distances are in mean free paths. The point source O is at a height H m.f.p. above a plane interface, below which there is a non-scattering medium. The point A at which the radiation is measured lies on this plane, and N is the foot of the normal from O . Let χ be the azimuthal angle of plane OAB , measured from the plane OAN .

If μ lies between $\{1 - (H/D)^2\}^{\frac{1}{2}}$ and 1, B lies in the scattering medium whatever the value of χ , and the dose from such directions is the same as if the scattering medium extended to infinity. For μ between $-\{1 - (H/D)^2\}^{\frac{1}{2}}$ and -1 , B lies inside the vacuum for all values of χ , and A receives no scattered radiation from such directions. If μ lies between $\pm \{1 - (H/D)^2\}^{\frac{1}{2}}$, radiation arrives only if χ lies between χ_0 and $2\pi - \chi_0$, where

$$\cos \chi_0 = H(D^2 - H^2)^{-\frac{1}{2}} \cot \theta. \quad (15)$$

In this region of μ , therefore, the integrand of the scattered dose in (11) must be multiplied by $1 - \chi_0/\pi$.

A 4. Point source in a spherical hole in an infinite medium

Let the spherical hole have a radius Y mean free paths. The source is at the centre. We seek the dose at points outside the hole. The unscattered radiation now contributes to the standard dose $e^{-(D-Y)}/4\pi r^2$. The modification to the scattered radiation depends on μ . Write

$$\mu_0 = \{1 - (Y/D)^2\}^{\frac{1}{2}}. \quad (16)$$

Then if $\mu \leq \mu_0$, of the path of the once-scattered radiation arriving at A a distance Y mean free paths is not in scattering material and the previous integrand, as in (11),

must be multiplied by e^Y . If $\mu_0 \leq \mu \leq 1$, the line AB in figure 2 intersects the surface of the hole at distances given by $x = \mu \pm (\mu^2 - \mu_0^2)^{\frac{1}{2}}$. If x lies between these values, the scattering point B lies inside the hole and there is no scattering. If $x < \mu - (\mu^2 - \mu_0^2)^{\frac{1}{2}}$ the integrand is multiplied by e^Y . If $x > \mu + (\mu^2 - \mu_0^2)^{\frac{1}{2}}$ the path in scattering medium is less by $Y + 2D(\mu^2 - \mu_0^2)^{\frac{1}{2}}$ mean free paths. Hence the standard dose, on the modified first-order method, is

$$\text{S.D.} = \frac{e^{-(D-Y)}}{4\pi r^2} \left[1 + \frac{2\pi D e^D 1.243 r_0^2}{\sigma_0 \sigma_a(\gamma)} \int_{\gamma/(1+2\gamma)}^{\gamma} E_1^{-2} f(\cos \phi, \gamma) dE_1 \right. \\ \left. \times \int_0^{\infty} F y e^{-D(x+y)} dx / |1 - x\mu| \right], \quad (17)$$

where $F(x, \mu) = 1$ if $\mu \leq \mu_0$ and if $\mu_0 \leq \mu \leq 1$ with $x < \mu - (\mu^2 - \mu_0^2)^{\frac{1}{2}}$;
 $= 0$ if $\mu_0 < \mu \leq 1$ and $\mu - (\mu^2 - \mu_0^2)^{\frac{1}{2}} \leq x \leq \mu + (\mu^2 - \mu_0^2)^{\frac{1}{2}}$
 $= e^{2D(\mu^2 - \mu_0^2)^{\frac{1}{2}}}$ if $\mu_0 < \mu \leq 1$ and $x > \mu + (\mu^2 - \mu_0^2)^{\frac{1}{2}}$.

It is found that the distribution of the scattered radiation over E_1 is scarcely altered by the presence of the hole in the medium.

The dose inside the hole could be evaluated without difficulty, but differs very little from $1/4\pi r^2$ at normal values of γ , say 1 to 5. The deviation is greatest at the surface of the hole, where scattered radiation with deflexions as small as 90° can contribute to the dose.

APPENDIX B. THE SOLUTION OF PROBLEMS WITH SPHERICAL SYMMETRY BY THE 'MODIFIED UPPER LIMIT' PROCEDURE

B1. Point source in a homogeneous medium

We write the standard dose at distance r , equivalent to D mean free paths, as $e^{-\alpha D}/4\pi r^2$, where α is a number depending on γ and D . In the 'modified upper limit' procedure we start with the standard dose due to unscattered and once-scattered radiation, namely (9):

$$\text{S.D.} = \frac{e^{-D}}{4\pi r^2} \left[1 + \frac{2\pi D e^D}{\sigma_0 \sigma_a(\gamma)} \int_{\mu=-1}^1 \int_{x=0}^{\infty} y^{-2} e^{-Dy - D\sigma_1 x/\sigma_0} \sigma_a(E_1) f(\cos \phi, \gamma) dx d\mu \right]. \quad (9)$$

Here $y^{-2} e^{-Dy}$ arises from the intensity of unscattered radiation at the scattering point B (figure 1). Actually the radiation at B gives a standard dose $e^{-\alpha Dy}/4\pi (yr)^2$, where α refers to the distance Dy . In the 'modified upper limit' we assume that all this radiation has the same value of $h\nu/mc^2$ and the same direction as the primary radiation at B , and that α can be given the value appropriate to distance D rather than to Dy . The effect is to replace e^{-Dy} in (9) by $e^{-\alpha Dy}$. We may also write for the whole standard dose $e^{-\alpha D}/4\pi r^2$, giving finally

$$e^{D(1-\alpha)} - 1 = \frac{2\pi D e^D}{\sigma_0 \sigma_a(\gamma)} \int_{\gamma/(1+2\gamma)}^{\gamma} \sigma_a(E_1) f(\cos \phi, \gamma) E_1^{-2} dE_1 \int_0^{\infty} y e^{-D(\alpha y + \sigma_1 x/\sigma_0)} dx / |1 - x\mu|. \quad (18)$$

This can be solved numerically by using trial values of α . Computation proceeds exactly as was explained in §A1, except that the analytical integration for $E_1 = \gamma/(1+2\gamma)$ now gives

$$\int_0^\infty y e^{-D(\alpha y + \sigma_1 x/\sigma_0)} dx / |1 - x\mu| = (e^{-D\alpha} + e^{-D\sigma_1/\sigma_0}) / D(\alpha + \sigma_1/\sigma_0).$$

In penetration through a plane slab of thickness D mean free paths we have, corresponding to equation (18), the equation (44) of Cave *et al.* (1950). In the plane case the integral on the right involves only a single integration over E_1 , and it can be shown that the integrand coincides with that of (18) when $E_1 = \gamma$. This is the region which gives the maximum contribution to the integral, and it is to be expected, therefore, that α will be much the same in the plane and spherical problems. The difference in α is in fact of order 0.01 at distances of a few mean free paths.

B2. Spherical source in homogeneous medium

The argument of §A2 applies. Equation (14) can be used with $s.D.(r, D) = e^{-\alpha D}/4\pi r^2$.

B3. Point source in a spherical hole in an infinite medium

We write the standard dose as $e^{-\alpha(D-F)}/4\pi r^2$, at points outside the hole. Proceeding in §A4, we find that we must replace (18) by

$$e^{(1-\alpha)D} - e^{(1-\alpha)F} = \frac{2\pi D e^D}{\sigma_0 \sigma_\alpha(\gamma)} \int_{\gamma/(1+2\gamma)}^\gamma \sigma_\alpha(E_1) f(\cos \phi, \gamma) E_1^{-2} dE_1 \\ \times \int_0^\infty G y e^{-D(\alpha y + \sigma_1 x/\sigma_0)} dx / |1 - x\mu|, \quad (19)$$

where $G(x, \mu) = 1$ if $\mu \leq \mu_0$ and if $\mu_0 \leq \mu \leq 1$ with $x < \mu - (\mu^2 - \mu_0^2)^{\frac{1}{2}}$;
 $= 0$ if $\mu_0 < \mu \leq 1$ and $\mu - (\mu^2 - \mu_0^2)^{\frac{1}{2}} \leq x \leq \mu + (\mu^2 - \mu_0^2)^{\frac{1}{2}}$;
 $= \exp[2D(\mu^2 - \mu_0^2)^{\frac{1}{2}} \sigma_1/\sigma_0]$ if $\mu_0 < \mu \leq 1$ and $x > \mu + (\mu^2 - \mu_0^2)^{\frac{1}{2}}$.

REFERENCES

- Cave, L., Corner, J. & Liston, R. H. A. 1950 *Proc. Roy. Soc. A*, **204**, 223 (part I).
 Corner, J. & Liston, R. H. A. 1950 *Proc. Roy. Soc. A*, **204**, 323 (part II).

Interferometric studies in diffusion

I. Determination of concentration distributions

BY C. ROBINSON

Research Laboratory, Courtaulds Ltd., Maidenhead, Berks

(Communicated by A. H. Wilson, F.R.S.—Received 4 July 1950)

[Plates 9 and 10]

An interferometer technique for the quantitative study of swelling and diffusion in transparent high polymers and a method of determining the concentration distribution from the fringe system have been developed. The concentration—distance curves in the system chloroform-stretched cellulose acetate have been determined at two different temperatures with a view to estimating activation energies over a range of concentrations. The curves are found to consist of two almost linear portions with a considerable and abrupt change of gradient taking place over a small range of concentration. A corresponding sudden change of gradient has been observed in other systems.

Considerable experimental difficulties were met in determining the refractive index and density curves for the two component systems, which were required by the method. Eventually a refractive index curve was constructed with the help of data obtained by several approaches.

Many biological and industrial processes are controlled essentially by diffusion in polymer systems for which the diffusion coefficients are greatly concentration dependent. Such processes cannot be completely understood so long as even the order of magnitude of the concentration dependence remains unknown. Although concentration dependences have been reported by several workers who have measured the diffusion coefficients for different concentration ranges, there are few cases where the diffusion—concentration coefficient relationship proper has been determined for concentrated polymer systems. Thus Hartley (1946) (when discussing the part played by variable diffusion coefficients in determining the behaviour of polymers when swelling and shrinking) was able to quote only the work of King (1945) on the permeability of keratin to water as an example of such results. Since then, apart from the work of Crank & Park (1949), which will be referred to later, there seems to have been only the paper by Rouse (1947), who used a steady-state method similar to that of King to study the diffusion of water in nylon and polythene.

The failure to determine the dependence on concentration of such diffusion coefficients has probably been due to the lack of satisfactory experimental methods, since certain difficulties arise both in the observation of concentration distribution and in the subsequent mathematical analysis, especially as the near-solid end of the concentration range is approached. In this and a subsequent paper, an interferometer method for the determination of diffusion coefficients in polymers will be described which, at least in the most favourable cases, is suitable for studying the whole concentration range from solid polymer to pure penetrant.

Essentially the method consists in determining concentration gradients by observing the interference pattern produced when unidirectional diffusion takes place within a specimen clamped between silvered plates. In a previous paper,

Robinson (1946), while discussing the orientation of polymer molecules which accompanies unidirectional diffusion into a polymer sheet whose swelling is restricted to one plane, has described an optical method for determining the distribution of double refraction which arises under such conditions. The interferometric method described in what follows was originally developed in order to supplement these results with data for the concentration distribution under the same experimental conditions, but the results obtained by the method are of more general interest. The development of the technique is illustrated by experiments with the system chloroform—cellulose acetate, but it should be applicable to any transparent polymer provided that the intrinsic double refraction is not too high compared to the refractive index difference between polymer and penetrant.

In diffusion experiments where the swelling is restricted to one plane, Hartley (1949) has shown that the rate of entry of the penetrant is greatly dependent on the orientation of the polymer molecules, penetration at right angles to the direction of stretch being much faster than parallel to this direction. In order to investigate this anisotropic penetration more fully, highly oriented sheet was used for the diffusion experiments. It was also expected that by using cellulose acetate with a definite degree of orientation more reproducible results would be obtained than with the ordinary sheet with its variable uniplanar orientation. The measurements reported in part I of this paper are limited to diffusion of chloroform at right angles to the direction of stretch. Diffusion parallel to this direction will be discussed in part II (Crank & Robinson 1951). Similar experiments with other penetrants will also be discussed in part II. The experiments with chloroform used for establishing this method were carried out with stretched cellulose acetate conditioned to a suitable uniform concentration of chloroform. This gave more rapid and more reproducible results than were obtainable with unconditioned, dry, cellulose acetate and avoided certain complications, including a possible dependence on the method of clamping (which seems to arise when using the unconditioned material) which has not at present been fully investigated.

While this work was in progress, a method of determining diffusion coefficient—concentration relationships from simple absorption experiments was developed by Crank & Park (1949) in this laboratory. In this method the rate of uptake of penetrant by polymer sheet is measured at various partial pressures of the former, and the resulting absorption—time curves are then analyzed mathematically (Crank & Henry 1949*a, b*) to obtain diffusion coefficient—concentration relationships. The limitations of both this and the interferometric method have not yet been fully explored, but it would seem that the two methods are to some extent complementary. Thus the absorption technique has advantages arising out of its simplicity and is independent of a knowledge of the density—concentration and refractive index—concentration relationships. It also appears to have advantages for very low concentrations of penetrant, while, on the other hand, the interferometric technique readily gives results for those higher concentrations of penetrant where the polymer flows under gravity. The interferometric method is particularly suitable for studying sharp transitions of diffusion behaviour, such as those which appear in the neighbourhood of the sharp boundaries often visible even under the microscope.

EXPERIMENTAL

The experimental procedure for determining the concentration distribution from which the diffusion coefficient—concentration relationship can be calculated is, briefly, as follows. A piece of polymer sheet is clamped between two almost parallel pieces of silvered glass, the method of clamping being such that it allows the wedge angle to be adjusted to a suitable small value. On immersion in the penetrant diffusion takes place and a concentration distribution is set up and, when observed in the interferometer, a fringe system is seen. If the slight wedge angle is allowed for, the refractive index change between any two points can be calculated from the number of fringes between those two points. Then, if the refractive index of one point has been identified, the concentration—distance curve can be obtained, provided the refractive index—concentration relationship is separately determined. Since, in calculating diffusion coefficients, concentrations are best expressed in g./cm.^3 , the specific volume—concentration relation must also be determined. The determination of the refractive indices and specific volumes presented certain difficulties and will be described in the latter part of the paper. Since swelling and diffusion are restricted to one plane, the double refraction of the polymer system will vary with the distance diffused whether or not the polymer was originally oriented (Robinson 1946). Hence, in order to use the correct refractive index—concentration relationship for interpreting the fringe system, it is convenient to determine the double refraction—distance distribution and to observe the fringe system through polaroid so oriented that the electric vector of the emerging light is either parallel or perpendicular to the direction of the polymer molecules.

(a) *The interferometer*

As shown in figure 1 the optical system was essentially that already described by Ambrose (1948) as a suggested attachment to a standard microscope, except that in the present apparatus all the components were mounted on saddle stands on an optical bench. *A* is a mercury lamp, *F* a no. 77 Wratten filter, *L* a plano-convex

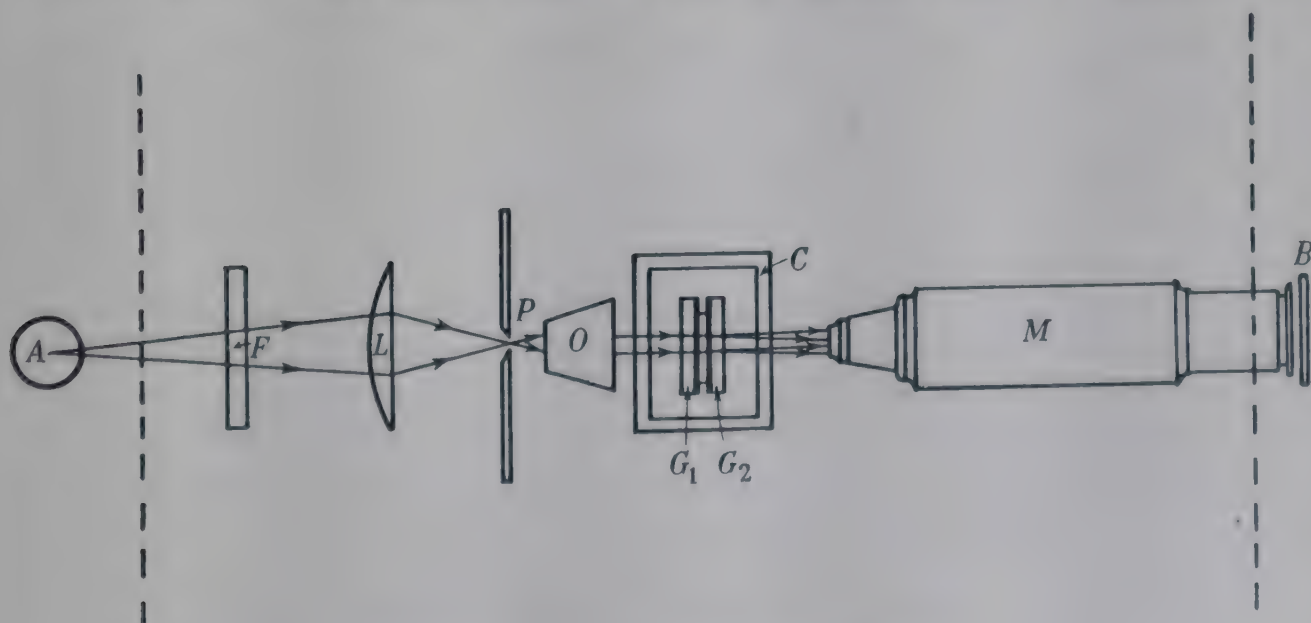


FIGURE 1. The interferometer optical system.

lens, P a pin-hole mounted, as described by Ambrose, so that it can be finely adjusted in all directions in a plane normal to the direction of the incident light, O a 16 mm. objective used as a collimator and fitted into a centring mount, C a glass cell as used in a Hilger absorptiometer, G_1 and G_2 two pieces of silvered plate glass between which the sheet of polymer was clamped, M the microscope and B a sheet of polaroid. The cell C was attached to a microscope mechanical stage which was mounted on the same saddle stand as the adjustable pin-hole and the collimator.

An optical system for producing Fabry-Perot fringes with silvered glass but without the fine adjustment for the pin-hole has been used by Berg (1938), Bunn (1949) and Humphreys-Owen (1949) for studying the concentration gradient of crystals immersed in solutions. The fine adjustment of the pin-hole introduced by Ambrose was essential for the success of the experiments to be described. By attaching full importance to it, sharp fringes could be obtained even with quite thick specimens. The explanation of how the adjustment of the pin-hole can be used to produce optical sharpness of the fringes has been given by Ambrose (private communication):

‘Although the fringe systems used in the experiments described in this paper have been formed by refractive index gradients, a qualitative picture of how they are produced can be seen by taking an example of an air wedge.

‘Figure 2*a* shows an air wedge formed between the flanking plates P_1 and P_2 in which the plane of the paper is a principal plane of the wedge. We are considering a special case when the rays incident upon the wedge lie in this plane. In a practical case they will also lie in directions outside this plane. A ray AB is incident normally upon P_1 . After successive multiple reflexions the rays AB , CD , EG , etc., will appear. The points from which pairs of these rays appear to diverge are denoted by F_1 , F_2 , etc. In order that any fringe system which is formed should be sharply focused the points F_1 , F_2 , etc., should lie as close together as possible.

‘In figure 2*b* the angle of the incident ray AB has been changed so that rays which are reflected within the wedge first travel in towards the corner of the wedge and then travel out once more. This is because the angles of incidence and reflexion are successively reduced as the rays approach the thin end of the wedge. This figure has been drawn on four times the scale of figure 2*a* in order to show the rays more clearly. Although an extra reflected ray AK has been included the points F_1 , F_2 , etc., are not more widely scattered than in figure 1*a*. For any given wedge angle there will be some angle of incidence for the ray AB which will give a minimum scatter of the points F_1 , F_2 , etc., and in this position the best fringes will be obtained. In the apparatus here used (figure 1) the direction of the ray AB is controlled by the sensitive pin-hole adjustment and it enables sharp interference fringes to be produced with thick films of polymer and steep refractive index gradients. The size of the pin-hole aperture is also critically related to the thickness of the specimen. In figures 2*a* and *b* the angles are, of course, grossly exaggerated and the effective wedge angle may in practice only correspond to a matter of minutes of arc, so that the correct phase relationships between rays after successive multiple reflexions can be maintained.

Searle (1946) has given a three-dimensional treatment to the problem of the formation of interference fringe systems by an air wedge, together with a number of illuminating experiments to illustrate the theory. No satisfactory explanation

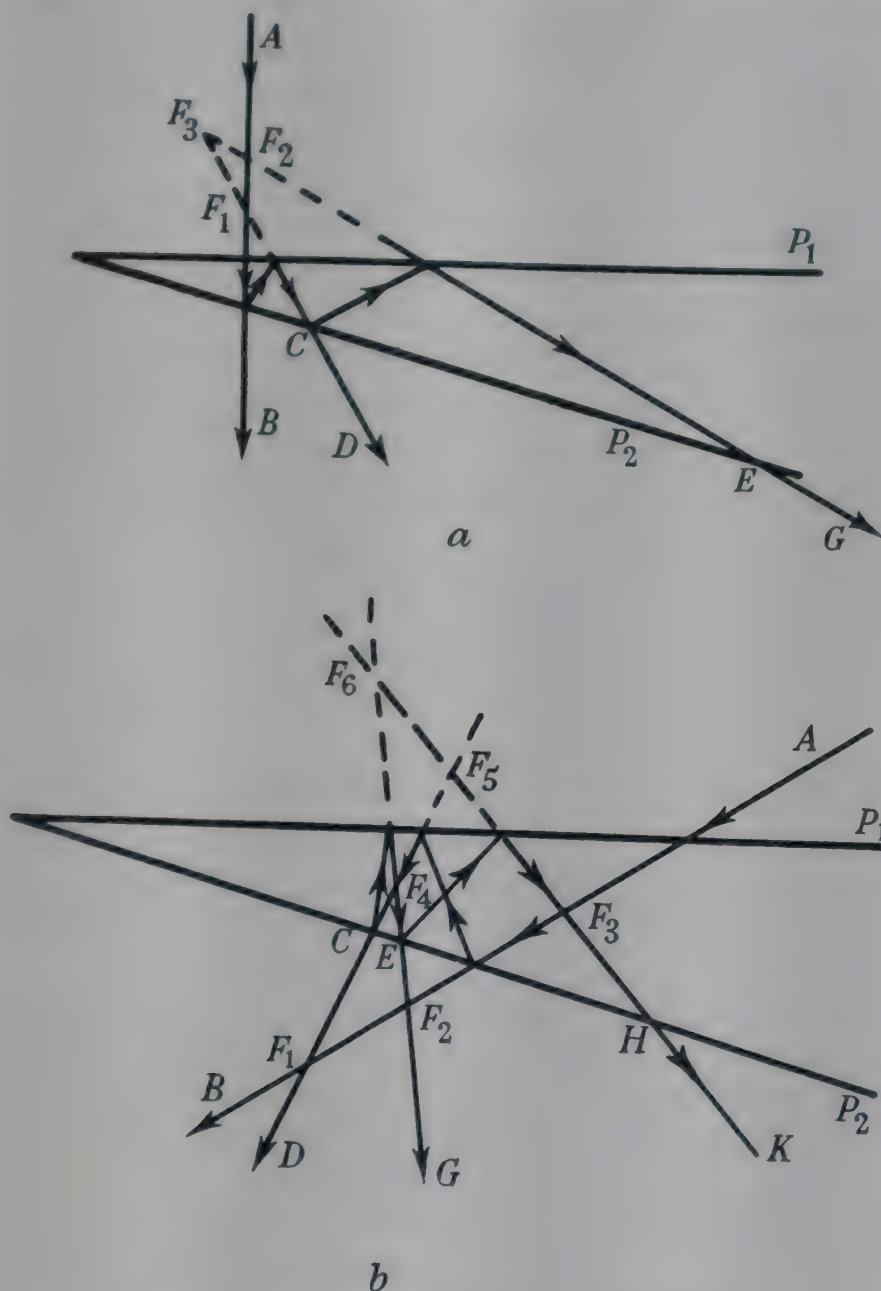


FIGURE 2. Path of light through air wedge.

has previously been given of the formation of this type of fringe system. This is because earlier treatments have only considered a two-dimensional case, which can never be realized in practice.'

In some of the diffusion experiments to be described where the refractive index changes vary abruptly at some point, it was found that the fringes on either side of this point could not be focused simultaneously in the two gradients corresponding to wedges of different 'focal length'. By suitably adjusting the pin-hole it was possible to produce optimal sharpness at any point in the fringe system, this being of most importance where the fringe spacing was small.

The whole apparatus between the dotted lines of figure 1 was enclosed in an air thermostat stirred by an electric fan and controlled by a contact thermometer. The apparatus was used in a room thermostatically controlled to within $\pm 1^\circ \text{C}$ of some temperature a few degrees lower than the temperature of the experiment. In these circumstances the temperature of the cell kept within $\frac{1}{20}^\circ \text{C}$. Focusing of the microscope and movement of the mechanical stage could be effected by controls passing through the walls of the thermostat while the adjustment of the pin-hole could be made by removing a small area of the cover.

Two types of mercury-vapour lamp were used, a G.E.C. Osira laboratory lamp and an Ediswan 'Escura' lamp. The former gave satisfactory narrow spectrum lines and a constant intensity. The latter provided a more intense source of light. The Escura lamp was used with the envelope removed. Even then it was necessary to run it at a reduced voltage to prevent the spectrum lines from being too broad. Under these conditions the intensity was not so constant as with the Osira lamp, which added to the difficulty of estimating the correct exposure.

(b) *The silvered glass*

The specimen was mounted between two pieces of $\frac{1}{8}$ in. silvered plate glass, $\frac{11}{16} \times \frac{13}{16}$ in. in area. The silvering was deposited by a vacuum evaporation method. The swelling of the cellulose acetate during the diffusion tended to strip the silver off the glass, small triangular holes developing in the silver, with their apexes pointing in the direction of swelling. These appeared in the neighbourhood of the middle boundary (to be described below) and increased in number and size until it was impossible to read the fringes in this region. It was found that the stripping could be avoided by depositing a very thin film of chromium on the glass before applying the silver.

Although increasingly sharp fringes are obtained by increasing the thickness of the silver deposit, it was necessary to compromise by using silvered plates of sufficiently high transmission in order that the exposure should be short enough when photographing the moving fringes. Deposits giving 2 to 20 % transmission were used, according to the nature of the experiment.

The silvered glass could be stored in a vacuum desiccator for weeks without serious loss of reflectivity. Good plate glass was used which did not show more than two fringes in the microscope field. It was not necessary to use glass which was more optically flat.

(c) *Mounting and experimental procedure*

The arrangement finally adopted for clamping the specimen between the pieces of glass is shown in figure 3. The clamp consisted of two H-shaped pieces of stainless steel, held together by two threaded studs and nuts. The wedge angle could be adjusted as required by tightening or loosening the nuts.

To mount the specimen, the strip of cellulose acetate was removed from the conditioning bottle and a rectangular piece about 10×5 mm. was cut off. A satisfactory straight edge parallel to the direction of stretch was obtained by using an ordinary pair of scissors. The piece was placed as quickly as possible

between the two pieces of glass (the silvered surfaces being in contact with the cellulose acetate) and the clamp tightened until Newton's rings had disappeared, so ensuring that no air was present between the glass and the cellulose acetate. The clamp was then adjusted so as to give a small, or negligible, wedge angle—that is, so that only a few fringes were seen in the interferometer field of view, making, preferably, a right angle with the edge of the acetate across which diffusion was to be observed. As the fringes were clearly visible in the laboratory fluorescent lighting, this adjustment could be made before placing in the interferometer cell. The difficulty of adjusting the wedge angle varied considerably with the conditioning concentration and the flatness of the specimen. During the mounting and adjusting some penetrant is lost from the edge of the material and a concentration gradient appears. The specimen, still mounted in the clamp, was therefore replaced in the conditioning bottle until no disturbance of the fringes at the edge could be detected.

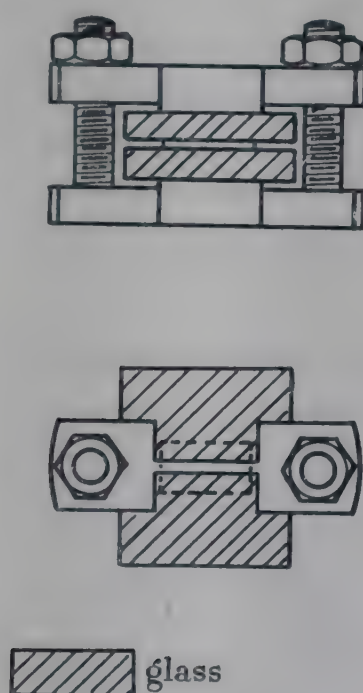


FIGURE 3. Clamping arrangement for glass plates.

Before beginning an experiment, the glass cell of the interferometer was filled with penetrant and brought to the temperature at which the measurement was to be made. The cellulose acetate, clamped between the glass plates, was then removed from the conditioning bottle and immediately immersed. A photograph was taken as soon as possible to record the initial wedge angle. In the most successful experiments there was no appreciable change in this angle throughout the period when observations were made. If the diffusion was allowed to go too far, softening of the acetate near the centre of the clamp caused a big change in the wedge angle, and further observations were worthless.

The distance travelled by the moving fringes was proportional to the square root of the time. It follows that in order to take early photographs the exposure must be sufficiently short. Ilford HP 3 plates were used and the camera bellows extension

was kept small so as to shorten the exposure. When using a $1\frac{1}{2}$ inch objective and illuminating a 1 in. diameter circle of the negative, the shortest satisfactory exposure was about 30 sec. (It was found that the exposure could be reduced to 2 sec. when using a 16 mm. cinematograph camera on account of the small size of the picture. In this way, therefore, earlier pictures could be obtained.) Enlargements (5 diameters) were made on contrast bromide paper. Measurements of the fringe spacings were sometimes made on the negative using a Hilger measuring microscope, but often it was found more convenient, and sufficiently accurate, to make measurements on the enlarged print, as in this way it was easier to see the whole field and avoid any region in which local blemishes occur in the cellulose acetate and where irregularities in the fringes are consequently found.

(d) *Measurement of thickness*

The thickness of the cellulose acetate was determined by measuring the total thickness of the glass plates and the acetate at the end of the experiment by means of a micrometer and then subtracting the thickness of the plates. It was important that the measurement should be made before the diffusion had gone far enough to cause any wedge angle change (as indicated by the fringes). This method was not as accurate as would have been wished. The thickness of the plates made it impossible to obtain a more accurate reading by using a microscope with a high-power objective.

INTERPRETATION OF FRINGE SYSTEM

Figure 4, plate 9, shows chloroform (on the left) diffusing into cellulose acetate conditioned to contain 20.3 % chloroform (percentage, here and throughout the paper, refers to grams of chloroform per 100 g. of mixture). The wedge-angle fringes in the chloroform run parallel to those in the undisturbed cellulose acetate and give no indication of any cellulose acetate going into solution. Reading from left to right, we next come to several very close fringes representing the outward swelling edge of the cellulose acetate. These are presumably due to the light passing through a meniscus formed by the liquid-gel interface and hence a varying thickness of the material.* Following these are a series of widely spaced fringes and then

* There are two main reasons for attributing these fringes to a meniscus. First, if early and late photographs of the same experiment are compared, the space occupied by the meniscus fringes is found not to have increased, although the spacing of fringes on the polymer side of the meniscus may have increased several fold. Secondly, after diffusion has proceeded for a considerable time, the number of these fringes is approximately the number which would be expected if the refractive index had changed from the value for fully swollen acetate to that for pure chloroform. This may be explained as follows: At a sufficiently late stage of the diffusion, the concentration throughout the meniscus may be considered uniform and equal to that of the fully swollen material. Then, on the left of the meniscus (figure 4) there will be a thickness l of pure chloroform, and on the right the same thickness, of fully swollen acetate. At any intermediate point there will be a certain thickness x of pure chloroform and a corresponding thickness $l-x$ of fully swollen cellulose acetate. It follows that the number of fringes between the two sides of the meniscus will be the same as would be given by a continuous change of concentration from fully swollen cellulose acetate to pure chloroform. The number of fringes would be the same, it may be added, with either a concave or convex meniscus, but their distribution would be different.

a quite abrupt change of the spacing as a region of closer fringes is reached. If fringe number is plotted against distance (as in figure 5, which is from a similar experiment)* two nearly straight lines of different slope are obtained, the transition region from one slope to the other covering not more than three fringes and possibly less. This transition corresponds to the boundary between isotropic and anisotropic regions observed by Hartley (1949) with the microscope and referred to in his paper as a 'middle' boundary. On the right of the closely spaced fringes, one can be seen which doubles back to form a sharp peak pointing upwards. This is owing to there being a maximum in the refractive index—concentration curve for the cellulose acetate-chloroform mixtures, so that on either side of this peak (in a case where the wedge angle is negligible) the refractive index is decreasing with

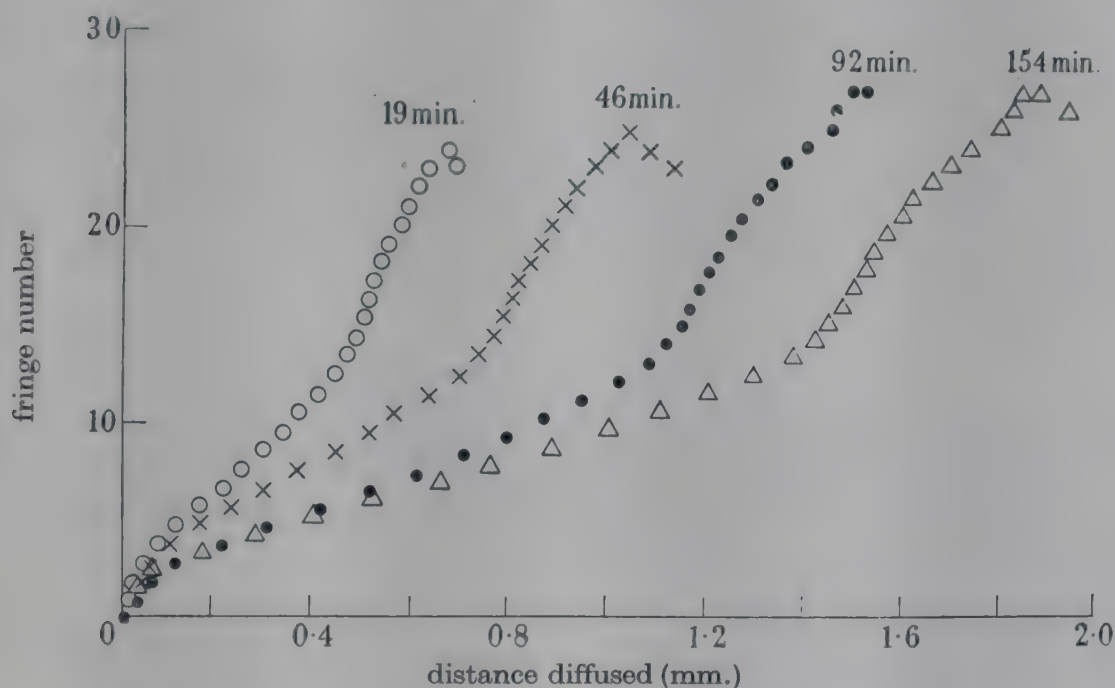


FIGURE 5. Fringe number—distance distribution at successive times. Experiment 44. Thickness 0.195 mm. chloroform entering from left. Initial concentration 21.7 % chloroform.

increasing distance from the peak. Finally this fringe can be seen curving gradually into the wedge-angle fringes of the undisturbed cellulose acetate. In cases where the wedge angle is appreciable and of suitable sign, a second, but more rounded peak will be seen pointing in the opposite direction to the refractive index peak. Here the change of optical path due to the wedge angle is first compensated by the change in the opposite direction due to the concentration gradient.

Where the wedge angle is small enough to be neglected it is possible to determine the refractive index at any point by counting the fringes between it and some other point of known refractive index (e.g. the boundary of the liquid penetrant). Then the refractive index difference between the two points is given by $n\lambda/2l$, where n is the number of fringes, λ the wave-length of the light used and l the thickness of the specimen. When the wedge angle is not negligible a correction must be applied to

* To show the several points discussed more clearly, it has been necessary to use data from different experiments in preparing figures 4, 5 and 6.

the value of n used. A method of making this correction will be described in part II. Where the fringe correction for the total distance over which diffusion has taken place was not more than ± 3 fringes, the general formula used in part II takes a simpler form. This leads to the following method which is sufficiently accurate for most of the experiments referred to here and has the advantage of being independent of the somewhat unsatisfactory thickness measurements. After the position of the fringes had been plotted, two points of known refractive index could be identified. Counting from the liquid end, the first fringe after the meniscus fringes (i.e. those fringes which do not lie on the first linear part of the curve) corresponds to the refractive index of the fully swollen material, while the bending back of a fringe identifies the position of maximum refractive index, the value of which can be read from the refractive index-concentration curve. Then, if n is the number of observed fringes between these reference points, the increment $\Delta\mu$ of refractive index per fringe is given by $\Delta\mu = (\mu_{\text{max.}} - \mu_{\text{fully swollen}})/n$.

When the difference in slope of the two linear parts of the fringe number—distance curve (as in the cellulose acetate—acetone system to be described in part II) is great, this method will not be sufficiently accurate even when the wedge angle is small, because then the effect of wedge angle will be relatively greater in the flatter portion of the curve.

CONCENTRATION—DISTANCE CURVES

To obtain the concentration distributions, the refractive index—distance relationships were calculated as explained above and then converted to concentration—distance relationships by using the refractive index—concentration and specific

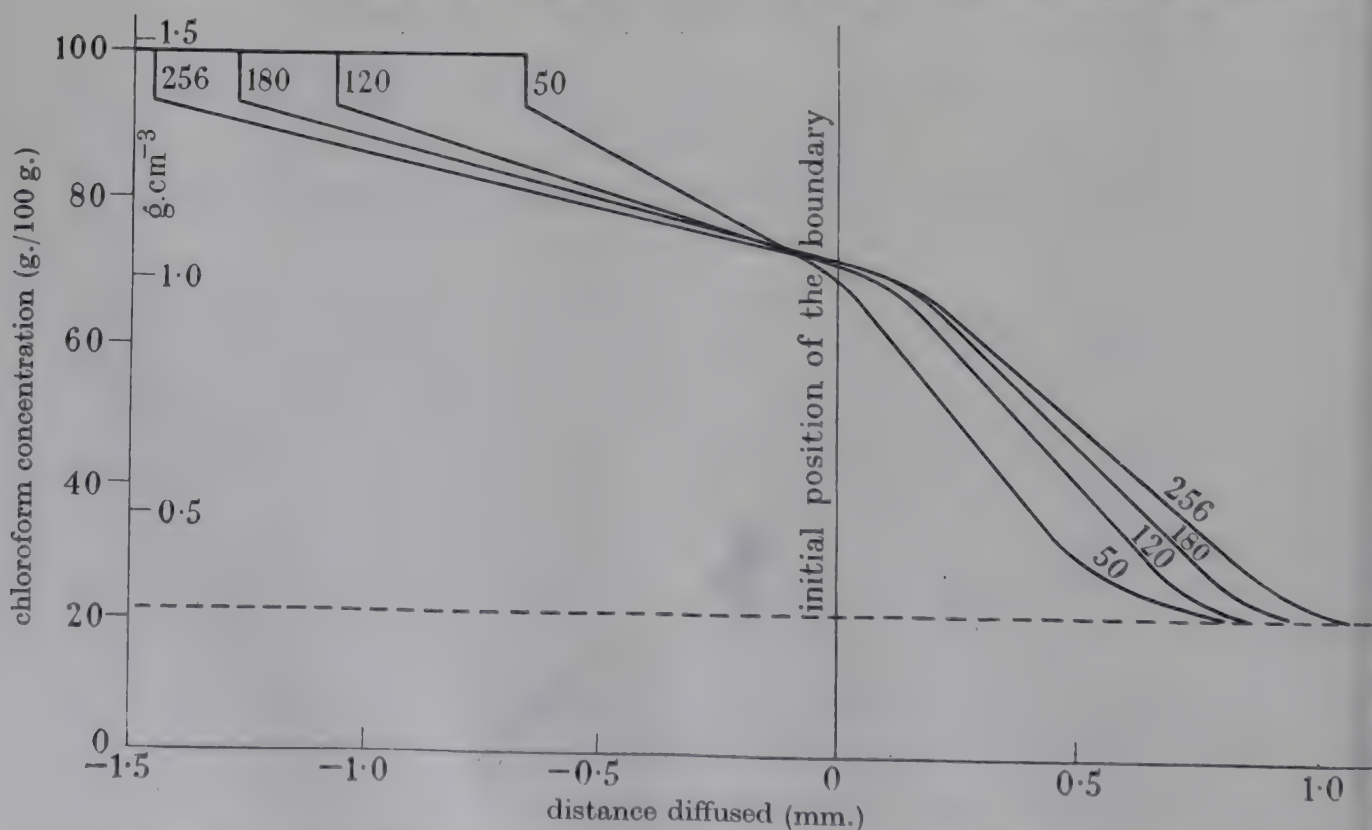


FIGURE 6. Concentration—distance distributions. Experiment 41. Thickness 0.20 mm. Initial concentration 22.3 % chloroform. (Numbers on curves denote time in minutes.)

volume—concentration curves of figures 10 and 12. The particular methods of obtaining the last two relationships, which presented some difficulty, will be described in detail in the later part of this paper.

In figure 6 concentration—distance curves are shown for one experiment at 25° C. Four photographs were taken at suitable intervals of time and a curve calculated from each photograph. The vertical part of each curve at the highest penetrant concentrations shows the discontinuity in passing from the liquid chloroform to the fully swollen polymer. The rest of the curves consist of two almost linear parts making a considerable angle with one another and joined together by a short transition region.

REPRODUCIBILITY AND THE SQUARE ROOT RELATIONSHIP

In calculating the diffusion coefficients from the concentration—distance curves by the method which will be described in part II, the assumption is made that the rate of

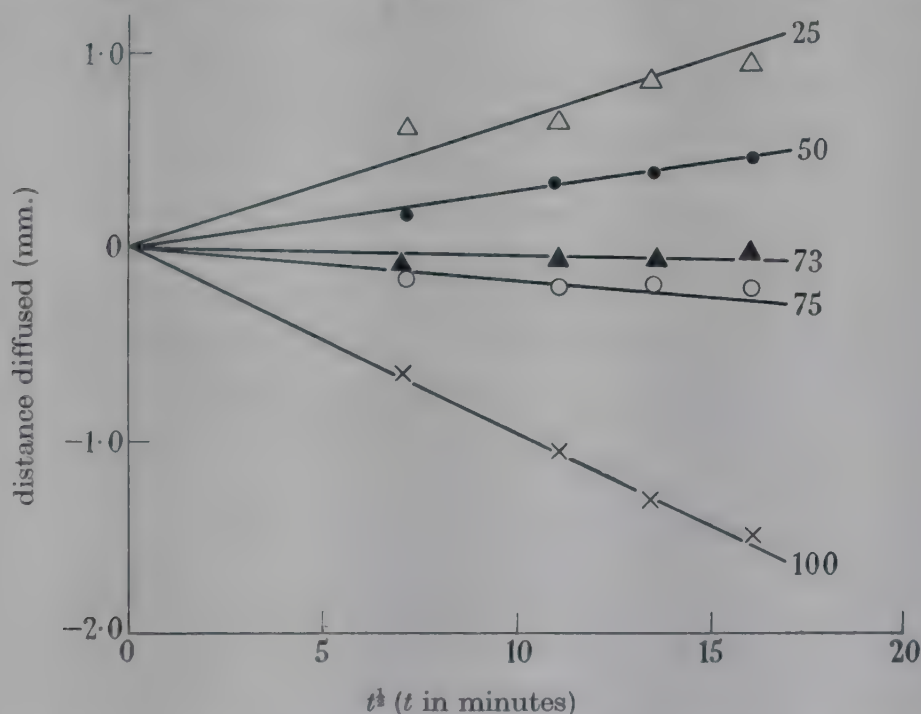


FIGURE 7. Distance diffused—square root of time. Experiment 41. The numbers are concentrations of chloroform expressed in g./100 g.

displacement of a given concentration is proportional to the square root of the time. That this assumption is experimentally justified will be seen from figure 7 where distance diffused (measured from the initial position of the boundary between penetrant and polymer) is plotted against the square root of the time and almost straight lines are obtained for each of the several concentrations. The departure from linearity does not seem to be greater than would be expected from the errors of the method. Similar results were obtained from other experiments with this system.

Figure 7 may also be used as evidence of reproducibility within an experiment. The reproducibility of results obtained between replicate experiments will be seen from the diffusion coefficient—concentration curves in part II (figures 2 and 3) for

25 and 40° C. In selecting experiments for publication only those have been rejected which were expected, for some understood reason, to have an abnormally large experimental error, such as where the wedge angle had collapsed, air films had not been eliminated or unreadable fringes occurred.

PREPARATION OF SPECIMENS

(a) *Method of stretching*

The cellulose acetate used was prepared from commercial sheet (0.25 mm. thick, referred to as foil in a previous communication (Robinson 1946)) which had been stretched and deplasticized as described below.

To obtain a high degree of orientation in cellulose acetate sheet, it must be stretched when containing some material which will allow sufficient stretching without appreciable relaxation of the orientation produced, and which can be removed after the treatment without reducing the degree of orientation. It follows that the material chosen must be a low molecular weight substance which can diffuse rapidly in and out of the cellulose acetate with as little swelling and shrinking as possible. The problem was investigated at some length in this laboratory by Hartley (unpublished research) who found that a high degree of orientation could be obtained without producing opacity in the sheet by using the hot vapour of methyl alcohol. The procedure adopted was as follows: A brass clamp was attached to each end of a strip of the cellulose acetate sheet (*not* deplasticized), 3.5 cm. wide and 10 cm. in length between the clamps. This was suspended vertically in the centre of a glass cylindrical double jacket and its temperature raised by a stream of air, heated to nearly 100° C. The vapour of boiling methyl alcohol was then introduced and passed for 20 min. with the air turned off, and then the strip was stretched by hand in a series of short sharp pulls until the length was increased by 2.6 times. (If elongated more than this it generally becomes opaque.) The alcohol vapour was immediately displaced by cold air and the strip left clamped at its increased length for some minutes while cooling. Finally, it was removed from the cylinder, given successive baths of cold methyl alcohol until no further plasticizer was removed, placed in an oven at 100° C for at least 24 hr. and then stored in a desiccator. Before being used for the conditioning and diffusion experiments, the strip was again dried in the oven. The temperature of the oven was not sufficient to reduce the double refraction of the sheet.

All the pieces used in the experiments had a double refraction between the limits of 1.2 to 1.3 orders as determined with a graduated quartz wedge, and a thickness of from 0.175 to 0.185 mm. when dry.

(b) *Method of conditioning*

The piece of cellulose acetate to be conditioned was hung from the ground-glass stopper of a bottle containing chloroform, diluted with redistilled butyl phthalate to give the required partial vapour pressure. Pure chloroform tends to decompose giving phosgene and hydrochloric acid. Methyl alcohol, which is a satisfactory stabilizer, could not be used on account of its high vapour pressure, but 2 % capryl

alcohol was found to be satisfactory, no decomposition being detected after several weeks. A vessel containing phosphorus pentoxide as a desiccant was also placed in the conditioning bottle, which was immersed in a thermostat until the weight of the strip had become sensibly constant. The weight was determined by rapidly placing the strip in a small weighing bottle and weighing on an aperiodic balance.

Above a certain concentration of chloroform, the cellulose acetate began to flow slowly and could no longer be safely suspended from a hook, so for concentrations above 50 % it was supported in the conditioning bottle in a glass vessel with small holes in the bottom to facilitate circulation of the conditioning vapour.

Cellulose acetate will absorb several units per cent of water from the atmosphere and the presence of this water was found greatly to increase the rate of penetration of the chloroform. Thus chloroform will advance a millimetre in a few hours into cellulose acetate which has been allowed to take up moisture, while penetration into oven-dried cellulose acetate is too slow to be measured in this time. It was therefore essential to exclude water from the system in order to obtain reproducible results. (Penetration into plasticized sheet is also rapid. It would seem that the effect of penetrant, water or plasticizer in speeding up the penetration is primarily the action of a diluent rather than anything more specific.)

AUXILIARY OBSERVATIONS

(a) *Double refraction*

The relationship between the double refraction distribution and the concentration distribution is needed for the interpretation of the fringe system and the middle boundary and also, as will be shown later, for the construction of refractive index—concentration curves.

The fringes produced by the undiluted stretched cellulose acetate are double when seen without polaroid in the optical system. This is because the stretched material acts like a negative uniaxial crystal (the direction of stretching corresponding to the optic axis). Light passing through the material emerges vibrating in two planes at right angles to one another and that in each plane produces a separate set of fringes. But if a piece of polaroid is placed in front of the eyepiece so that the electric vector of the transmitted light is either parallel, or at right angles to, the direction of stretch, only one set of fringes will be seen. Figure 8, plate 9, shows a photograph of chloroform diffusing into stretched cellulose acetate, the direction of diffusion being perpendicular to the direction of stretch. In both the top and bottom strip of the field of view a piece of polaroid was inserted, while there was no polaroid in the centre strip. The two pieces of polaroid were in crossed positions and so oriented that the fringe systems in the strips which they covered were formed by polarized light having the plane of vibration of its electric vector respectively parallel and normal to the direction of diffusion. By comparing the fringes in the different strips it can be seen how the double refraction of the cellulose acetate gradually falls from 1.2 orders to zero with increasing dilution and that the isotropic region extends from the middle boundary to the outer boundary which indicates the greatest extent of the swelling and dilution.

For all quantitative observations of the concentration distributions, photographs were always taken (as in figure 4) with a sheet of polaroid in front of the eyepiece. Although, with the type of polaroid used, this increased the exposure time fourfold, it was more convenient to use than a Nicol prism, which would introduce a smaller aperture. It was, of course, necessary to note which way the polaroid was oriented and then to use the appropriate refractive index—concentration curve.

To determine the change in double refraction with distance in a diffusion experiment a similar technique to that already described by Robinson (1946) for isotropic cellulose acetate and acetone was used. The cellulose acetate was mounted exactly as for an interferometer experiment except that the glass was not silvered. The diffusion was then observed in a polarization microscope using a Wright eyepiece fitted with a graduated quartz wedge and a first-order red selenite plate in the focal plane of the objective. (A second first-order red plate had to be added below the objective so that all the curve, from the highly oriented to the isotropic material, would be within the field of view.) The black line which is then seen, corresponding to the thickness of the quartz wedge which just compensates the two first-order red plates, is displaced where it crosses the refractive index gradient by an amount which is proportional at any point to the double refraction at that point. The slope of this 'quartz-wedge curve' indicates, therefore, the rate of decrease of double refraction with distance as we pass from highly oriented to isotropic cellulose acetate. Figure 9, plate 10, shows such a photograph taken after 60 min. It will be seen that the double refraction falls continuously with this highly stretched material. This is unlike the curves obtained for isotropic, or only slightly oriented, cellulose acetate using acetone (Robinson 1946) or other penetrants (Hartley 1949) where the double refraction rises to a maximum value because the orientation produced by swelling (stretching) causes a greater increase of double refraction in lower concentrations of penetrant than the decrease which arises from dilution.

By taking successive photographs of the same experiment and measuring the position of points on the quartz-wedge curves, it was possible to show that, within the accuracy of reading the curve, the rate of advance of any given double refraction value is proportional to the square root of the time. Now since, as has been seen, the interferometer results show that any given *concentration* advances proportionally to the square root of the time, it follows that any given value of the double refraction will correspond to the same concentration at whatever time it was observed and that the decrease in double refraction with concentration remained constant throughout the period when photographs were taken. (The constancy of the height of the peak of the double refraction as it advances during the diffusion of acetone into unstretched cellulose acetate (Robinson 1946) may be looked upon as a special case of this relationship. The effect will be found either if there is no relaxation and the decrease in double refraction is directly due to dilution, or, if relaxation is fast compared to the shortest time interval after which an observation is made.)

For constructing the refractive index curve (figure 12) the decrease of double refraction was taken as linear, since inspection of the quartz-wedge curve shows

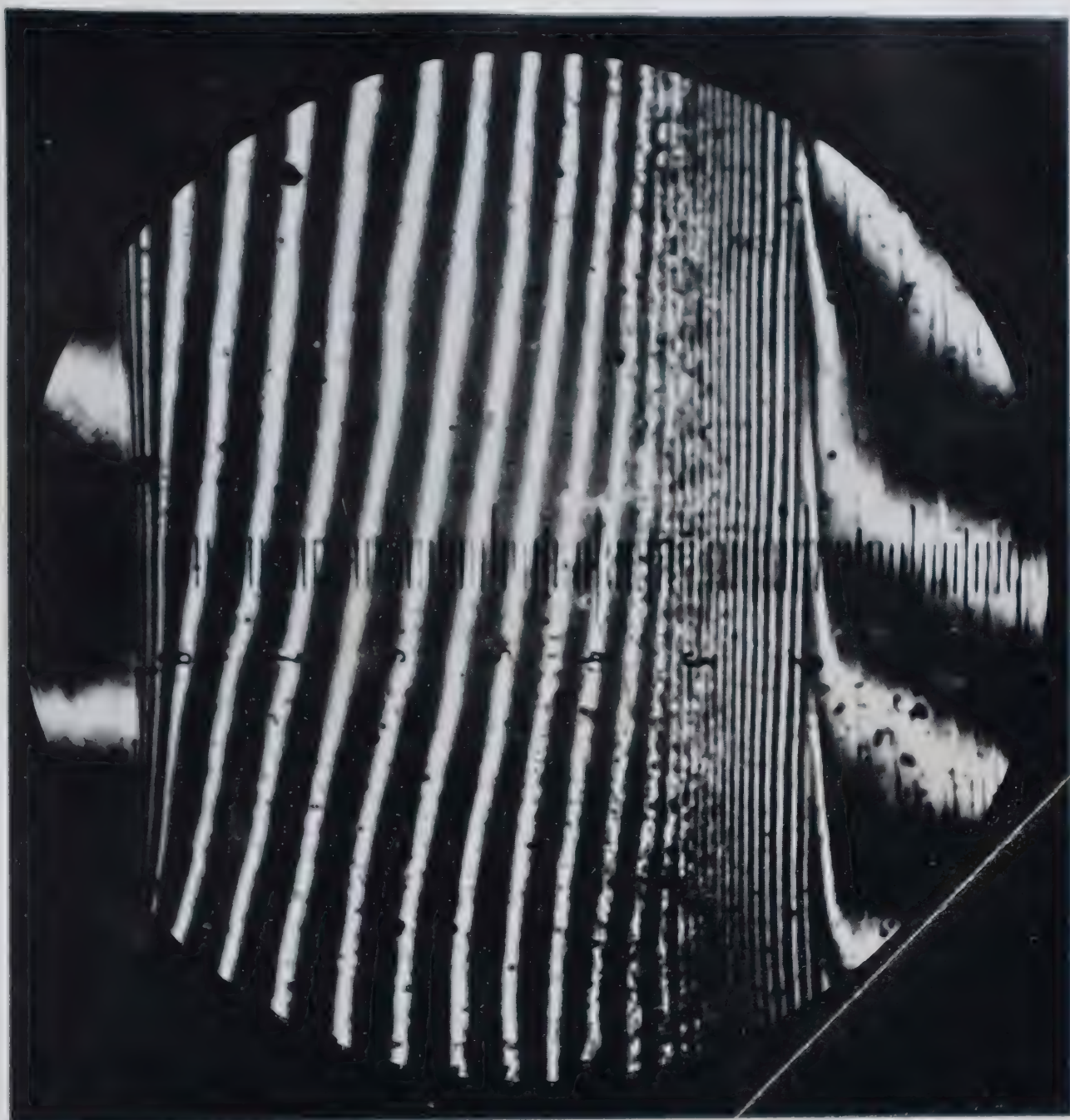


FIGURE 4. Experiment 129. Cellulose acetate conditioned to 20.3 % chloroform. Chloroform entering from left; 25° C; thickness 0.200 mm.; time 155 min.; magnification $\times 49$.

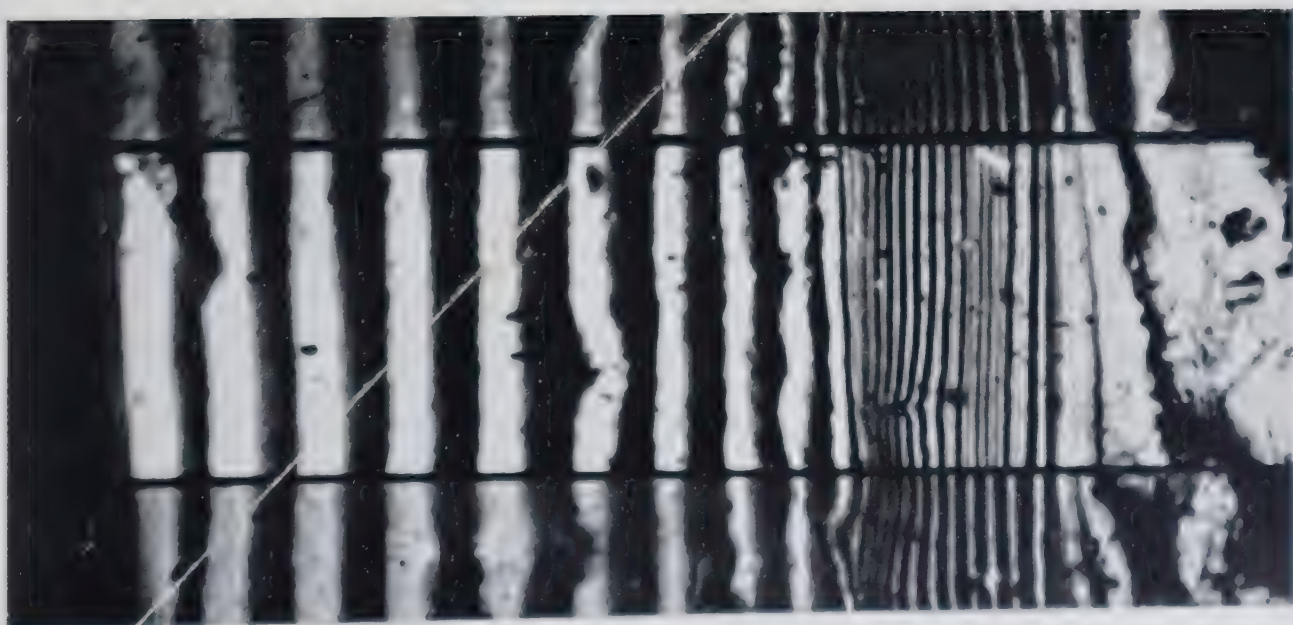


FIGURE 8. Experiment 91. Doubling of fringes in stretched cellulose acetate; chloroform entering from left; 25° C.

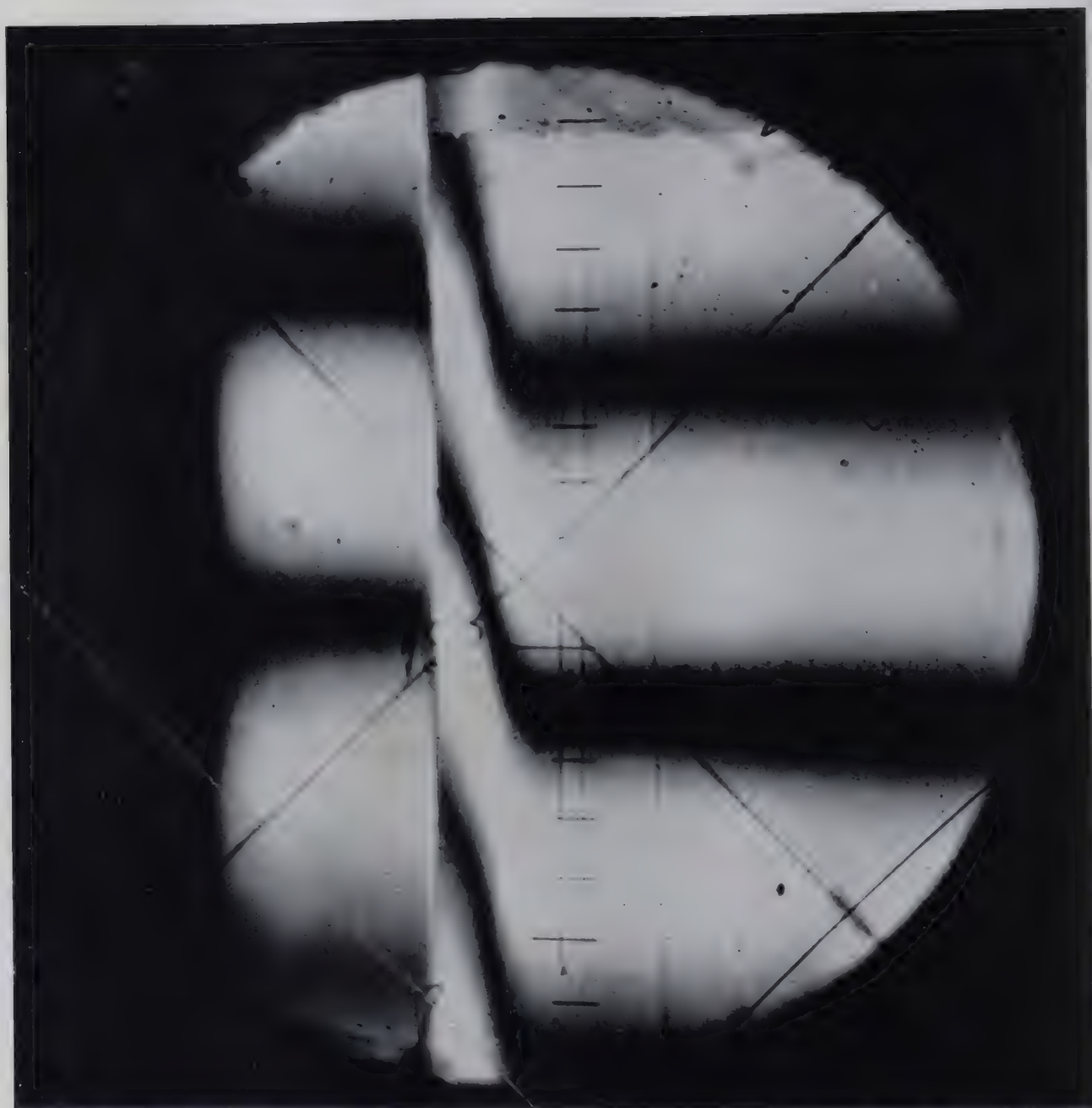


FIGURE 9. Experiment 46. Double refraction—distance curve formed with quartz wedge; stretched cellulose acetate conditioned to 21.3 % chloroform; chloroform entering from right; 25° C; thickness 0.200 mm.; time 60 min.; magnification $\times 22$.

that no great error in the refractive index is introduced by this simplification. Actually, as will be seen from figure 9, plate 10, the curve falls more steeply than this in two regions, but the value of the double refraction is never far from what it would have been if the fall had been linear throughout the anisotropic region.

By using this quartz-wedge technique for determining the double refraction distribution, the double refraction values are obtained which actually arise during a diffusion experiment, while if the double refraction had been determined for each concentration by an observation on a piece of cellulose acetate conditioned to that concentration, other values would have been obtained. This is because in the diffusion experiments swelling normal to the sheet is prevented, while a piece hanging freely in the conditioning bottle can increase in thickness with a consequent tendency to increase the orientation of molecules normal to the sheet.

(b) *Determination of specific volumes*

The method used for determining the density of the pieces of cellulose acetate conditioned with a volatile penetrant must be such that errors due to loss of penetrant by evaporation or mixing with an immersion liquid are not serious. It was found that the use of a density gradient column as described by Linderström-Lang, Jacobsen & Johansen (1938), and used for polymers by Boyer, Spencer & Wiley (1946), was suitable and convenient provided the penetrant concentration was not too high.

The density gradient column was prepared from mixtures of kerosene and carbon tetrachloride and calibrated by drops of zinc chloride and lead nitrate solutions of known density. A small piece of the cellulose acetate whose density was to be measured was cut off immediately after removal from the conditioning bottle and allowed to fall into the cylinder. The time of exposure to the atmosphere was not more than 6 sec. The pieces containing low concentrations of chloroform fell to a position of equilibrium in about 20 sec., coming to rest with the plane of the sheet horizontal, which allowed the position of their centres of gravity to be determined accurately with a cathetometer. The pieces containing a higher chloroform content inevitably lost some chloroform by diffusion into the cylinder while falling and, instead of remaining motionless at one height, began to rise in the column after having reached a maximum depth. The specific volumes shown in figure 10 correspond to these minimum positions without correction. It follows that for the higher chloroform contents the true values for the specific volumes are if anything slightly higher than those reported. Above 40 % the error became too large and no results have been given.

For concentrations between 32 and 74 % chloroform specific volumes were obtained by weighing a comparatively large strip (*c.* 1.5 g.), conditioning to the required concentration in chloroform vapour and weighing again after rapidly transferring it to a weighing bottle, and then (after a short reconditioning) weighing the strip, while suspended in chloroform, in a metal cage from the beam of an aperiodic balance. The cage arrangement was necessary to allow the sample to be weighed while immersed in a liquid having a greater density than itself. To obtain this last weight, successive weighings were made over 2 or 3 min. in a constant-

temperature room, during which time the chloroform content increased by diffusion into the strip. The weight was then plotted against the square root of the time and extrapolated to zero time by drawing a straight line. The correction was, of course, highest for the lowest chloroform concentrations.

The specific volume of the fully swollen material was determined by swelling a strip in liquid chloroform while suspended in the metal cage and then weighing while suspended in chloroform. The equilibrium swelling concentration was then determined by weighing the strip in a weighing bottle after gently mopping off surplus chloroform with filter paper. The error in the determination was small, owing to the very high liquid content.

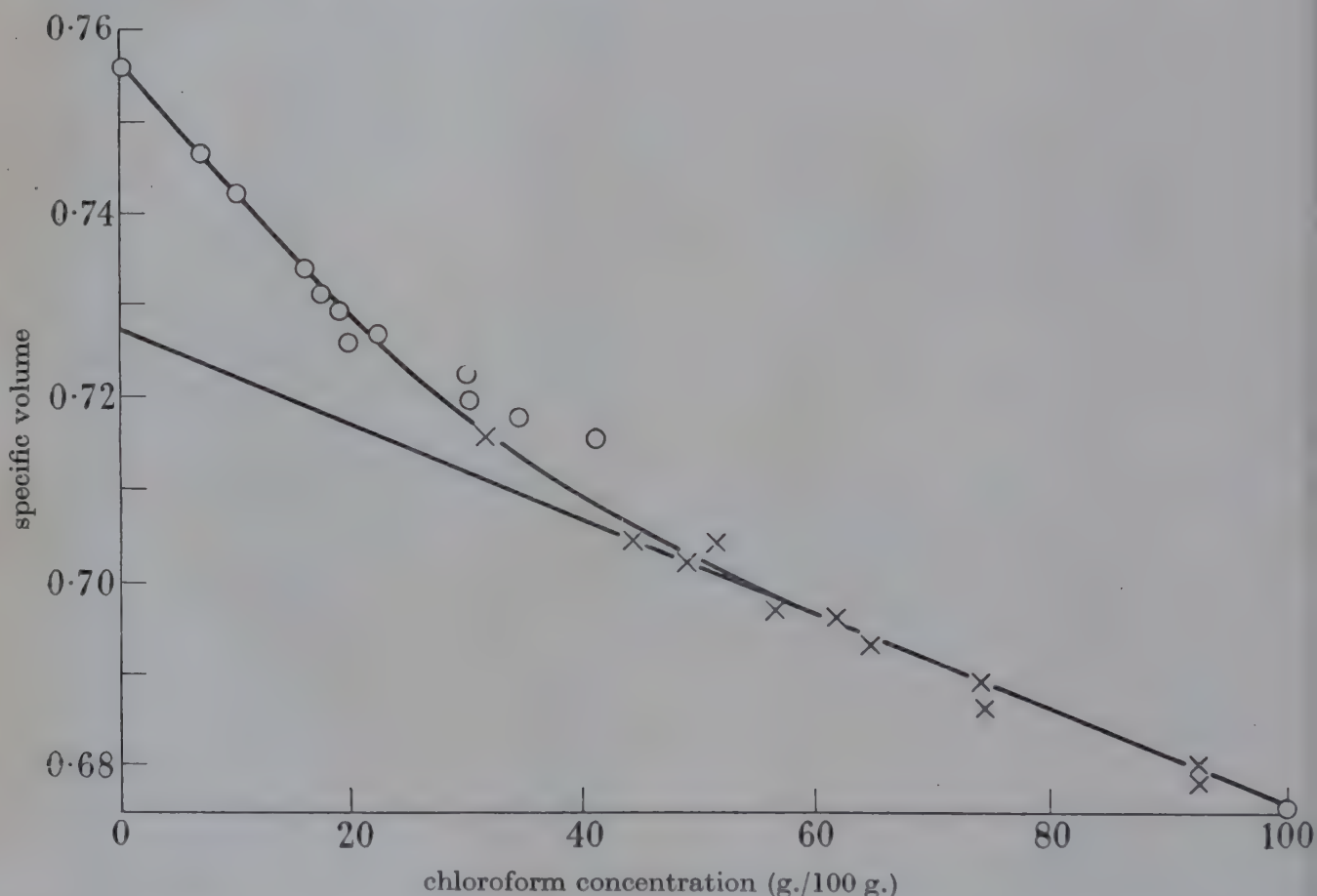


FIGURE 10. Specific volume of cellulose acetate—chloroform mixtures at 25° C.
O density column; × immersion weighings.

The specific volume—concentration curve was completed by drawing a straight line through the values obtained for concentrations above 40 % and the specific volume of pure chloroform and then extending the smooth curve which could be drawn through the values below 25 % until it met this straight line tangentially. In drawing the curve it is assumed that all the density column determinations above 25 % gave too high a specific volume.

(c) *Dependence of specific volume on free space*

The shape of the specific volume—concentration curve is consistent with the conception that the specific volume of the conditioned material may be looked upon as the sum of the partial specific volumes of the two components and a

certain volume of free space, as in P. H. Hermans's (1946) treatment of viscose and water.

If V_c is the volume occupied by 1 g. of cellulose acetate when fully swollen, V_p the volume of 1 g. of the penetrant, and V the volume of 1 g. of the mixture containing a g. of penetrant, we have

$$V = \epsilon + aV_p + (1-a)V_c, \quad (1)$$

where ϵ is the free space per gram of mixture still remaining at concentration a . This free space is defined as the free space which is not available to the penetrant at concentration a , but which becomes available at sufficiently high concentrations, where

$$V = aV_p + (1-a)V_c. \quad (2)$$

Assuming that all the volume change on mixing is due to the occupation of this free space, the contraction due to the mixing at concentration a is given by $\epsilon_0 - \epsilon$ (where ϵ_0 is the free space present in 1 g. of dry cellulose acetate) and for zero concentration of penetrant

$$V = \epsilon_0 + V_c. \quad (3)$$

By linearly extrapolating the straight part of the specific volume curve for 25° C to cut the specific volume axis, a value of 0.728 cm.³/g. is obtained for V_c and 0.028 cm.³/g. for ϵ_0 . The magnitude of this free space is of interest when penetration with little or no swelling is being considered in part II. The corresponding curve for viscose and water can be drawn from Hermans's carefully determined data. The curve has a very similar shape and gives a value of 0.036 ml./g. for ϵ_0 . In contrast, one may compare mixtures of chloroform and *n*-amyl alcohol, in which non-polymer system the specific volume varies linearly with the composition.

In equation (1) the volume of 1 g. of the conditioned cellulose acetate was expressed as the sum of its three components. Following Hermans's treatment for viscose and water mixtures by applying the law of Gladstone & Dale additively to the system, and remembering that $\mu - 1$ will be zero for free space, it is possible to calculate μ for all values of a if μ_p , V_p , μ_0 and V_0 have been determined experimentally. In this way we obtain a refractive index-concentration curve for the isotropic material, showing a refractive index maximum, similar to that shown in figure 11 for the experimentally determined values. The double refraction for the dry material can be obtained from the Rayleigh interferometer readings and the relative *fall* in the double refraction with distance diffused, from the quartz-wedge curve. It is therefore possible to construct the two refractive-index curves for light polarized at right angles and parallel to the long axes of the molecules or the direction of stretching. For this we used the treatment applied to cellulose by P. H. and J. J. Hermans (1946) who point out that for this material the two principal polarizabilities in a plane at right angles to the cellulose chain may be treated as if they were equal, in which case it can be shown that $\mu_{\parallel} - \mu = 2(\mu - \mu_{\perp})$, where $\mu_{\parallel}$ and $\mu_{\perp}$ are the refractive indices for light polarized parallel and at right angles to the direction of stretch and μ is the calculated refractive index of the isotropic material. Since, as will be explained in the next section, difficulty was encountered in directly determining the refractive indices of the chloroform-cellulose acetate mixtures, it was hoped that such calculated values could be used

in calculating the diffusion coefficients. It was found, however, that in experiments with cellulose acetate conditioned to a given concentration of chloroform, the number of fringes between the peak of the fringe system and the effectively undisturbed cellulose acetate was always more than would be expected from the calculated refractive index curve. With cellulose acetate conditioned to 21.8 % chloroform, there were, for example, four fringes in excess. Attempts to explain this discrepancy were unsuccessful. It was therefore necessary to use a curve obtained by some other method. The calculated results however confirmed that the curve exhibits a maximum and that its shape is similar to that shown in figure 11.

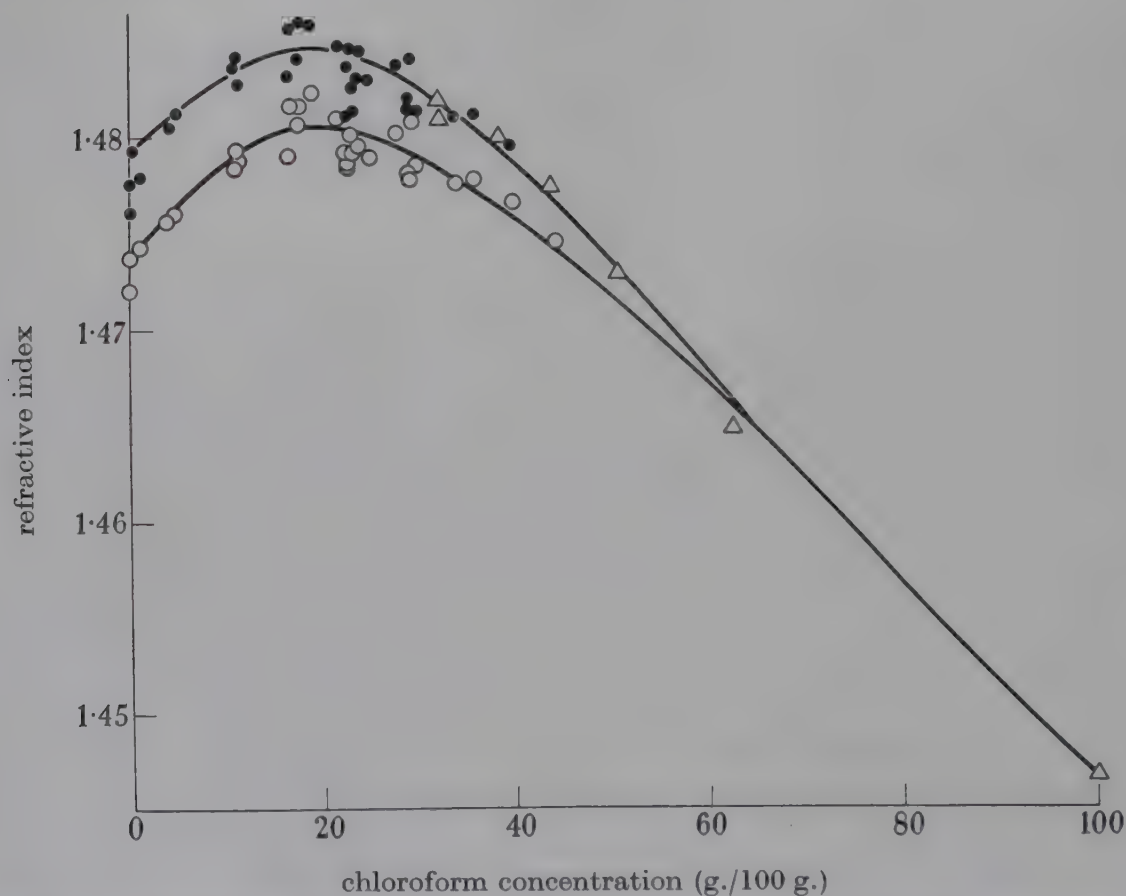


FIGURE 11. Refractive indices of cellulose acetate—chloroform mixtures at 25° C (directly determined). ● Rayleigh interferometer, electric vector perpendicular to direction of stretch; ○ Rayleigh interferometer, electric vector parallel to direction of stretch; △ Abbé refractometer measurements.

(d) *Refractive indices*

If a piece of cellulose acetate which has been brought to equilibrium in a certain partial pressure of penetrant is removed from the conditioning bottle to, for instance, a microscope slide, the refractive index in the immediate neighbourhood of the surface will be changed through loss of penetrant affecting both concentration and orientation. It follows that those methods of determining refractive index which depend on measuring the surface refractive index (e.g. Becke line method, Abbé refractometer, etc.) would introduce serious errors in determining the refractive index-concentration curve.

To get over this difficulty Dr Elliott suggested using a modified Rayleigh interferometer which he designed for the purpose. In this design, the slots covering the collimating lens were sufficiently small (2.0×0.4 mm.) to allow the piece of film covering them to be of uniform refractive index. To obtain a reading, a compensator is rotated until the two white fringes of a movable fringe system come exactly underneath the corresponding white fringes of the fixed reference system. The refractive index is calculated from successive readings in two immersion liquids of known and different refractive index. Not only must the liquids have suitable refractive indices, but they should be immiscible with the chloroform and not penetrate the cellulose acetate. Eventually the following two mixtures were adopted, which, for the time of immersion involved, were fairly satisfactory in this respect, partly on account of high viscosity: 1 part of lactic acid in 4 parts of glycerin and 1 part of citric acid in 2 parts of glycerin by weight. These approximate concentrations gave refractive indices of 1.4620 and 1.4740 at 25°C .

Unfortunately, except with the lowest concentrations of penetrant, what should have been the white fringes of the movable system were often somewhat coloured, evidently because of a difference in the dispersion of the specimen and the immersion liquid. In such cases it was difficult to decide which fringe to match and it was possible for an error of from 0.002 to 0.004 to arise, according to whether the wrong fringe was read in one, or both, liquids.

The Rayleigh interferometer technique could not be used for high penetrant concentrations since the material became too fluid for mounting. Some readings with the Abbé refractometer were therefore taken, but they were subject to serious errors through changes in surface concentration by exudation or evaporation. The Abbé refractometer was also used to determine the refractive indices of the chloroform and the immersion liquids.

The results shown in figure 11 are therefore subject to considerable errors and it was necessary to find some other method of estimating these values. They are, however, reported as showing the general nature of the curve with the pronounced maximum which was also obtained from the calculated results; the true position of the maximum may, however, as later considerations indicate be at a higher refractive index and a higher chloroform content. The values at the lowest concentrations should be fairly accurate. In drawing the curves of figure 11 for the refractive indices of light whose electric vector is respectively parallel and perpendicular to the direction of stretch, the middle boundary has been used as indicating the concentration below which the material becomes isotropic and where, therefore, the curves become coincident. The refractive index of this point was determined by observing the number of fringes between it and the outer swelling edge, and the mean thickness of the specimen. In drawing the curve the value of this refractive index was used, and the curve was assumed to be linear for concentrations above this point, an assumption which seemed unlikely to produce any serious errors.

The refractive index curve finally used for interpreting the results was obtained, following a suggestion of Mr Crank's, from direct observations of the fringe systems given by diffusion experiments with pieces of cellulose acetate conditioned to a range

of concentrations of chloroform (13.6 to 43.8 %). The maximum refractive index was determined from the number of fringes between the peak and the outward swelling edge, in the same way as was used in determining the refractive index of the middle boundary. The value for the maximum used in the graph (figure 12) is the average of the results obtained in this way from the several experiments on pieces conditioned to concentrations below the value giving the maximum. The refractive index corresponding to each conditioning concentration was then estimated by counting the fringes from the peak in the particular experiment to the undisturbed cellulose acetate and then subtracting the calculated difference in refractive index from the averaged value for the maximum refractive index.

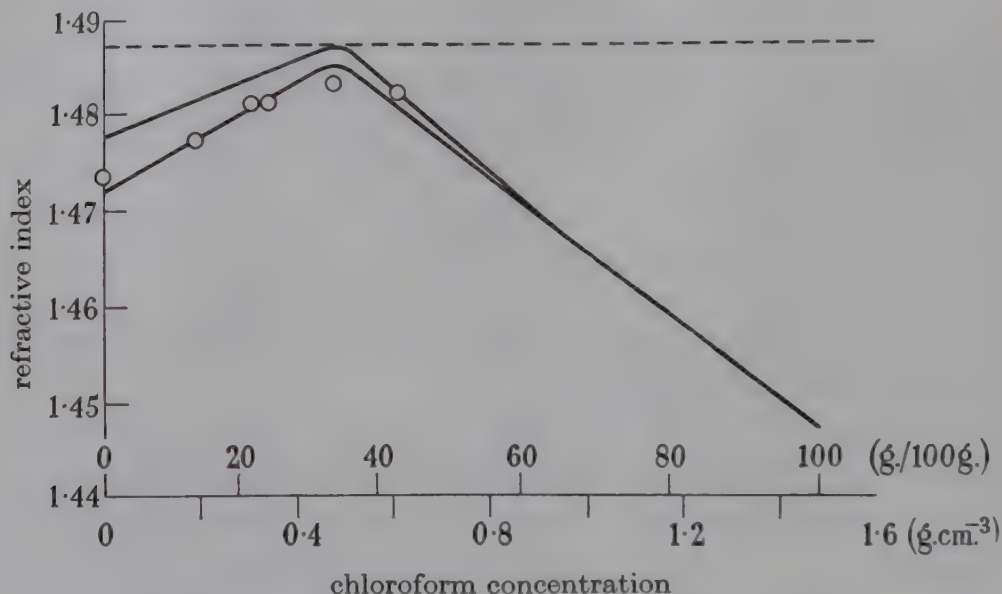


FIGURE 12. Refractive indices of cellulose acetate—chloroform mixtures at 25° C (diffusion experiments).

A smooth curve, consistent with the averaged maximum value was drawn through these points and those given by the measured refractive indices of cellulose acetate and chloroform. The point where the two curves became coincident was taken as being the same as in figure 11 and the curve was drawn as linear for concentrations above this point.

In calculating the diffusion coefficients at 25° C this curve was used together with the specific-volume curve of figure 10. A corresponding refractive-index curve was constructed, using the observed values at 40° C for the pure components and calculating intermediate points from the 25° C curve on the assumption that the temperature coefficient varies linearly with concentration. The refractive index of the peak so obtained agreed well with the calculated value obtained after counting fringes from the outer swelling edge in photographs of diffusion experiments at 40° C.

Thanks are due to my colleagues, Dr A. Elliott and Mr E. J. Ambrose, for advice and discussions concerning the optical problems involved and to Mr J. Goodwin for much painstaking technical assistance. I am also grateful to Dr G. S. Hartley for his stimulating suggestions in the early stages of the work.

REFERENCES

- Ambrose, E. J. 1948 *J. Sci. Instrum.* **24**, 134.
 Berg, W. F. 1938 *Proc. Roy. Soc. A*, **164**, 79.
 Boyer, R. F., Spencer, R. S. & Wiley, R. M. 1946 *J. Polymer Sci.* **1**, 249.
 Bunn, C. W. 1949 *Disc. Faraday Soc.* no. 5132.
 Crank, J. & Henry, M. E. 1949a *Trans. Faraday Soc.* **45**, 636.
 Crank, J. & Henry, M. E. 1949b *Trans. Faraday Soc.* **45**, 1119.
 Crank, J. & Park, G. S. 1949 *Trans. Faraday Soc.* **45**, 240.
 Crank, J. & Robinson, C. 1951 *Proc. Roy. Soc. A*, **204** (in the Press).
 Hartley, G. S. 1946 *Trans. Faraday Soc. B*, **42**, 6.
 Hartley, G. S. 1949 *Trans. Faraday Soc.* **45**, 820.
 Hermans, J. J. 1946 Appendix in J. J. Hermans' *Contribution to the physics of cellulose fibres*. Amsterdam: Elsevier Publishing Co.
 Hermans, P. H. 1946 *Contribution to the physics of cellulose fibres*. Amsterdam: Elsevier Publishing Co.
 Humphreys-Owen, S. P. F. 1949 *Proc. Roy. Soc. A*, **197**, 218.
 King, G. 1945 *Trans. Faraday Soc.* **41**, 479.
 Linderström-Lang, K., Jacobsen, O. & Johansen, G. 1938 *C.R. Lab. Carlsberg*, **23**, 17.
 Robinson, C. 1946 *Trans. Faraday Soc.* **42 B**, 12.
 Rouse, P. E. 1947 *J. Amer. Chem. Soc.* **69**, 1068.
 Searle, F. C. 1946 *Phil. Mag.* (vii), **37**, 361.

Surface energy and structure of crystal boundaries in metals

BY K. T. AUST AND B. CHALMERS

Department of Metallurgy, University of Toronto

(Communicated by E. N. da C. Andrade, F.R.S.—Received 15 July 1950)

[Plate 11]

The variation of boundary energy γ_R in lead with the difference of orientation θ was determined using the methods previously described for tin.

It was found that the energy decreases progressively as the difference of orientation decreases below about 16° . The curve relating energy with difference of orientation extrapolates to zero energy at zero angle. These results agree qualitatively with those of tin and offer further strong support for the 'transitional lattice' theory of the crystal boundary.

Straight-line plots of $\log \theta$ against γ_R/θ were obtained for lead and tin. These results are closely consistent with Shockley and Read's theoretical predictions based on a dislocation model of the transition boundary.

INTRODUCTION

A metal in its usual condition consists of an aggregate of small crystals; the cohesion between the crystals is strong and the neighbouring crystals exert powerful influences on each other's mechanical properties. In order to understand these matters, it is necessary to know the structure of the metal in the regions between the crystals that do not belong entirely to one crystal or to the other. In a previous paper (Aust & Chalmers 1950), evidence was presented which strongly supported the theory

that the crystal boundary is a region of transition from the lattice of one crystal to that of the other; the alternative theory, that of an amorphous 'cement', appears to be inconsistent with this evidence.

If it is recognized that the crystal boundary is a 'transitional lattice', it follows that each atom has a most probable position, which is determined by the minimum free-energy condition when the positions of all the atoms are considered. The position of each atom is, therefore, determinate and, in principle, calculable. There appears to have been no successful attempt to make this calculation, but an alternative approach consists in postulating a structure, calculating the properties of such a boundary and comparing the predictions with experimental data.

Bragg (1940) and Burgers (1940) pointed out that the transition from one orientation to any other can be achieved by introducing the appropriate array of dislocations. On this basis, Shockley & Read (1949) have calculated the variation of the specific energy of the boundary between cubic crystals whose difference of orientation is small. The results of Dunn & Lionetti (1949) are consistent with this theory but do not extend over a wide enough range to be regarded as providing strong support.

The present paper describes an investigation of the variation of boundary energy in lead with the difference of orientation, and discusses the interpretation of the results in terms of Shockley & Read's analysis; it is concluded that the results are closely consistent with their predictions.

2. EXPERIMENTAL METHODS

Lead of 99.999 % purity was used for the preparation of specimens consisting of three crystals of predetermined orientations. The technique was similar to that previously described for tin (Aust & Chalmers 1950). The seed crystal *A* (figure 1*a*) was oriented with $\langle 100 \rangle$ axes parallel and perpendicular to the length and plane of the specimen. Two rectangular 'legs' *B* and *C* were then grown from the seed *A* using a suitably shaped graphite boat with a removable insert at *D*. The difference in orientation of any required amount was introduced between the parts *B* and *C* by twisting *C* about its longitudinal axis by the required amount. Another seed crystal was then introduced at *E*; its orientation was such that a $\langle 100 \rangle$ axis was inclined at 45° to the plane of the boat. Three seed crystals, *B*, *E* and *C* were then present side by side and the specimen was grown from them. The immediate result was the production of two crystal boundaries (*PF* and *QF*), which, by virtue of the orientation of *E*, converge towards each other. This convergence occurs in lead only when the speed of growth is fairly high, i.e. not less than about 5 mm./min. After the boundaries *PF* and *QF* have met, a third boundary, *FR*, is formed (figure 1*b*).

The difference of orientation θ at the boundary *FR* was determined from back reflexion Laue X-ray photographs taken at points in the two crystals near the boundary. The orientation of each crystal is uncertain by $\pm 1^\circ$ owing to the 'macro-mosaic' effect that has been described previously (Chalmers 1949). The specimens were annealed at $315 \pm 2^\circ \text{C}$ for periods up to 100 hr., and the angles between the

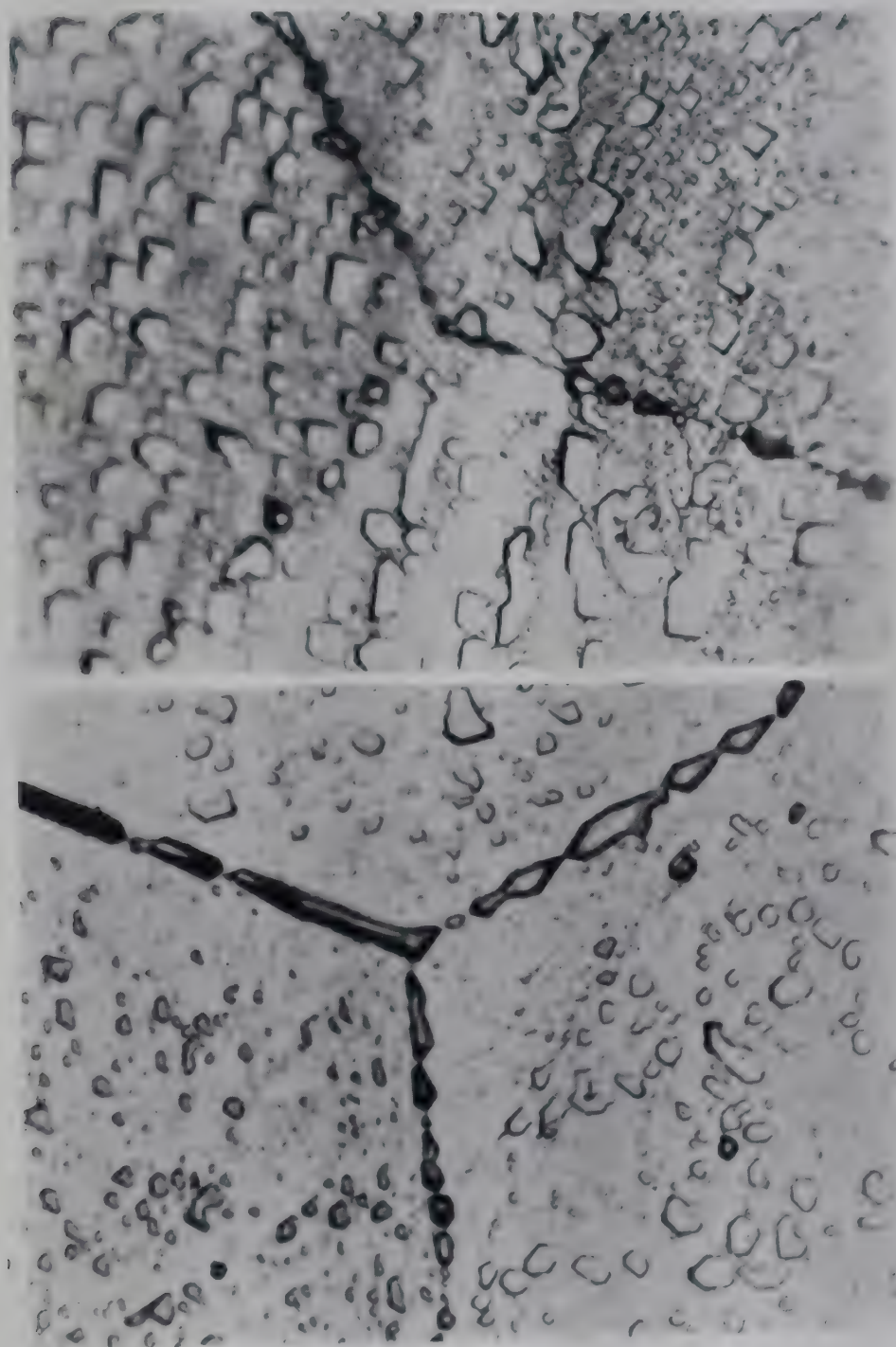


FIGURE 2.

three boundaries were measured after etching in a 10 % nitric acid solution. In order to avoid inaccuracy due to possible curvature of the boundaries, linear magnifications of 300 to 1000 times were used for this measurement. The accuracy of this measurement depended on the definition with which the boundaries could be observed. It was estimated from the consistency of repeated measurements of the same angles that the accuracy varied from $\pm 1^\circ$ in favourable cases to $\pm 3^\circ$ in the least satisfactory specimens. Typical photographs of the boundaries near the point of contact are shown in figure 2, plate 11. It was ascertained by successive observations after increasing periods of annealing that the angles had reached equilibrium values in

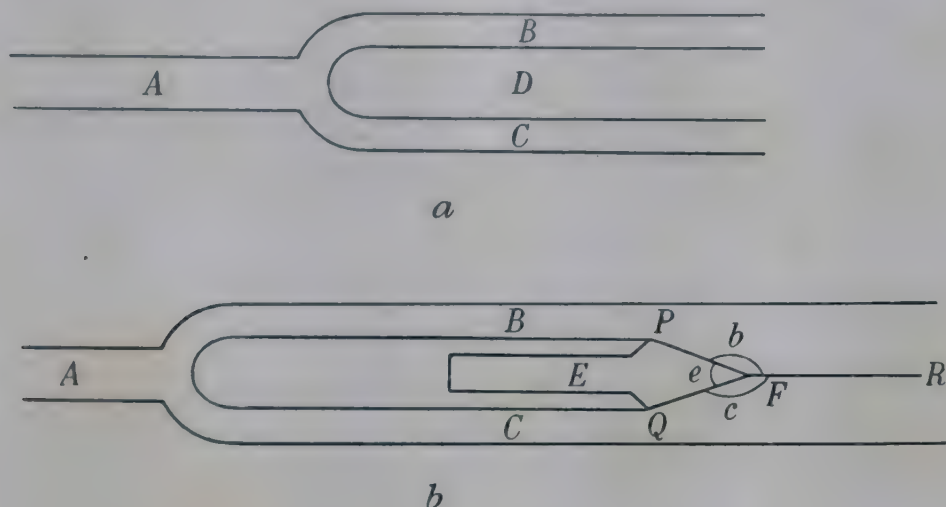


FIGURE 1.

40 to 100 hr. The angles e , b and c (figure 1 *b*) were recorded after equilibrium had been reached. They were used for calculating the relative free energies (γ_R) of the surfaces FR , FP and FQ respectively by means of the relationship

$$\frac{\gamma_{FR}}{\sin e} = \frac{\gamma_{FP}}{\sin c} = \frac{\gamma_{FQ}}{\sin b}. \quad (1)$$

It has been pointed out by Herring (1950) that this equation is only valid if the free energy is independent of the direction of the boundary in relation to its surroundings. A case in which equation (1) could not be used is when the boundary is between crystals with the twin relationship, as the boundary energy would then increase very sharply as it deviates from the twinning plane. In the present case, however, there is no evidence of any such singularity, and no attempt has been made to apply the more general treatment of Herring.

3. EXPERIMENTAL RESULTS

The equilibrium angles and the corresponding differences in orientation are shown in table 1. The relative free energy γ_R of the boundary FR is also given, on the basis of unit free energy of the boundary FP . The values of γ_R are calculated from equation (1) using unity as the value of γ_{FP} .

Figure 3 shows the variation of γ_R with θ , for lead, derived from table 1.

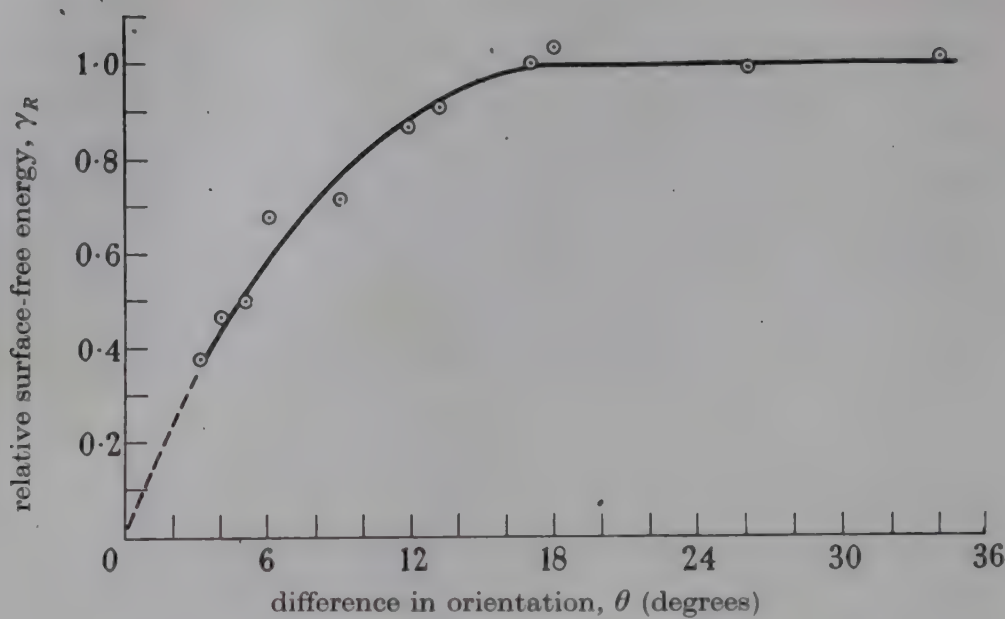


FIGURE 3.

4. DISCUSSION

It will be observed that a marked reduction in relative surface-free energy takes place when the difference in orientation drops below about 16° and that the decrease is progressive. The curve extrapolates to zero energy when $\theta = 0$, i.e. when the two crystals have the same orientation. The corresponding results for tin were shown in a previous publication (Aust & Chalmers 1950).

TABLE 1*

specimen number	orientation difference θ in degrees across <i>FR</i>	angles			relative surface-free energy γ_R
		<i>e</i>	<i>c</i>	<i>b</i>	
1	34	119	118	123	1.02
2	26	121	116	123	0.99
3	18	117	123	120	1.04
4	17	120	124	116	1.00
5	13	126	117	117	0.91
6	12	128	108	124	0.88
7	9	138	110	112	0.72
8	6	140	110	110	0.68
9	5	151	102	107	0.50
10	4	153	103	104	0.47
11	3	158	100	102	0.38

It will be evident that, qualitatively, the results for lead are similar to those for tin. Although there is some difference between the shapes of the curves, they both indicate that it is only for small angles of θ that γ_R shows any variation and that it then decreases progressively towards zero at $\theta = 0$. The results for lead, therefore, support the conclusion that was reached previously (Aust & Chalmers 1950) that the boundary cannot be amorphous and must, therefore, have a definite structure.

A possible structure of a crystal boundary as an array of dislocations has been discussed by Burgers (1940) and by Bragg (1940), and has been analyzed in some detail by Shockley & Read (1949). The proposed structure is represented diagrammatically in figure 4.

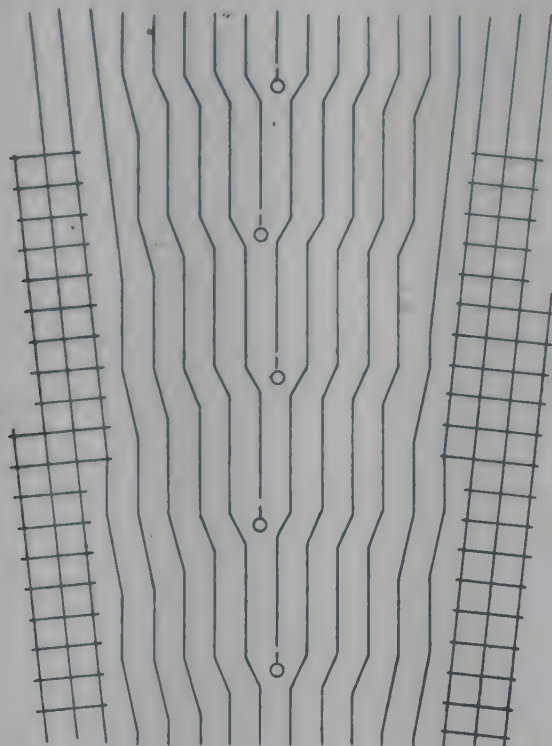


FIGURE 4. A transition surface between crystallites formed by a set of parallel edge dislocations.

Shockley (1949) states that, if we have two crystals with a small orientation difference, then a fit can be accomplished between these two across any arbitrary plane which joins them by putting in a prescribed distribution of dislocations. Rules for calculating the distribution can be made and crystal boundary energies for small orientation differences can be calculated on certain assumptions. Shockley & Read (1949) arrived at the following equation for the absolute interfacial energy γ_A , using the dislocation model of the boundary:

$$\gamma_A = \frac{ua}{4\pi(1-\delta)} \theta(A - \ln \theta)$$

for an isotropic cubic metal where

u = shear modulus,

δ = Poisson's ratio,

a = slip vector of the dislocation used,

A = a constant depending upon the energy near the dislocation where Hooke's Law fails,

θ = orientation difference where θ is small.

It follows that the experimental values of γ_A/θ should be related linearly with $\ln \theta$ with slope = $\frac{ua}{4\pi(1-\delta)}$. It also follows that there should be a linear relationship

between γ_R/θ and $\log \theta$. The experimental results for lead are plotted in this way in figure 5; it will be seen that a good straight line is obtained.

The following substitutions were made in the formula for γ_A in order to obtain the value of A which fits the experimental results for lead: $\delta = 0.45$, $a = 3.5 \times 10^{-8}$ cm., using for a the closest approach of atoms in lead; $u = 1 \times 10^{11}$ dynes per cm.², this

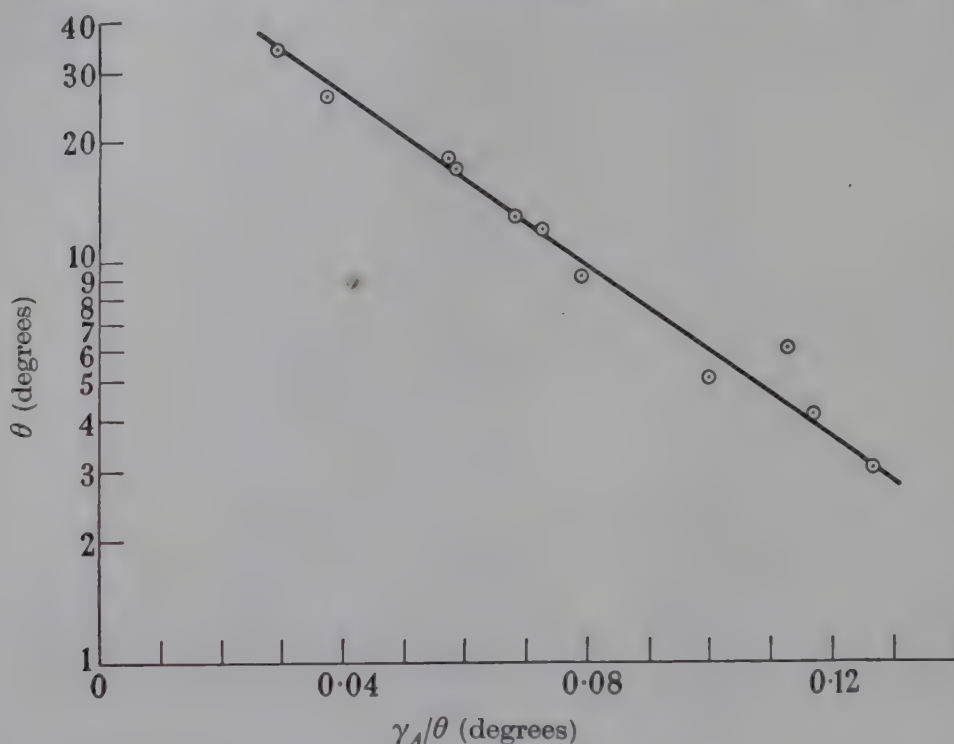


FIGURE 5.

being the average of the maximum and minimum values of shear modulus for a single crystal of lead (Boas 1947); experimental values of γ_R and θ such as $\theta = 34^\circ$, $\gamma_R = 1.02$ and $\theta = 3^\circ$ and $\gamma_R = 0.38$. These substitutions gave the result for lead that $A = +0.2$. The absolute energies γ_A of the boundaries in lead were then calculated using the formula for γ_A , the value of A as 0.2, and the experimental relative energies γ_R . These values are superimposed as points on the theoretical curve of absolute energy against orientation difference calculated only from the formula for γ_A and $A = +0.2$ (figure 6, curve 1).

It is concluded that the experimental results for lead are in close agreement with the predictions which Shockley & Read made on the basis of the dislocation model of the boundary. The most surprising feature of the results is that the relationship is linear for values of θ up to about 35° , whereas the theory is only expected to apply to small values of θ .

A similar analysis of the data for tin (Aust & Chalmers 1950) gave the following results. Figure 7 shows the relationship between γ_R/θ and $\log \theta$, and it is reasonable to apply the formula in the same way as for lead. Using $\mu = 1.5 \times 10^{11}$ dynes/cm.², $a = 3.0 \times 10^{-8}$ cm., $\delta = 0.33$ and the experimental figures, it was found that $A = -0.6$. The physical significance of the negative value of A is not understood, but it may be a consequence of applying the results of Shockley & Read's analysis for a cubic metal to the tetragonal structure of tin. Using this value, however, results are obtained which are shown in figure 6, curve 2.

Since Chalmers (1949) showed that the relative surface-free energies are constant for large values of θ , it is obvious, as Shockley & Read stated, that their calculations are only for small values of θ . This is also evident by the negative energy obtained if curve 2, figure 6, is extrapolated to larger values of θ .

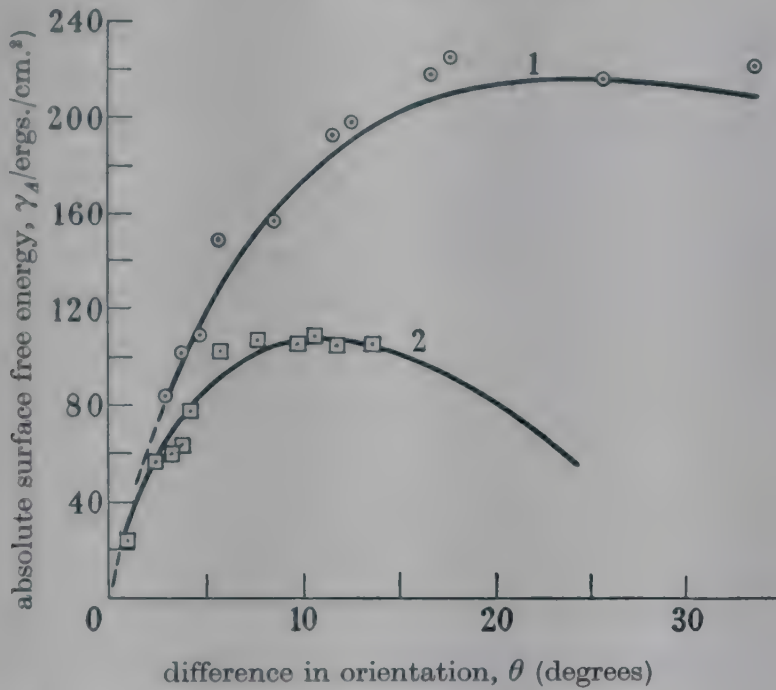


FIGURE 6. 1, theoretical curve for lead, $A = +0.2$; 2, theoretical curve for tin, $A = -0.6$; ○, □ experimental points.

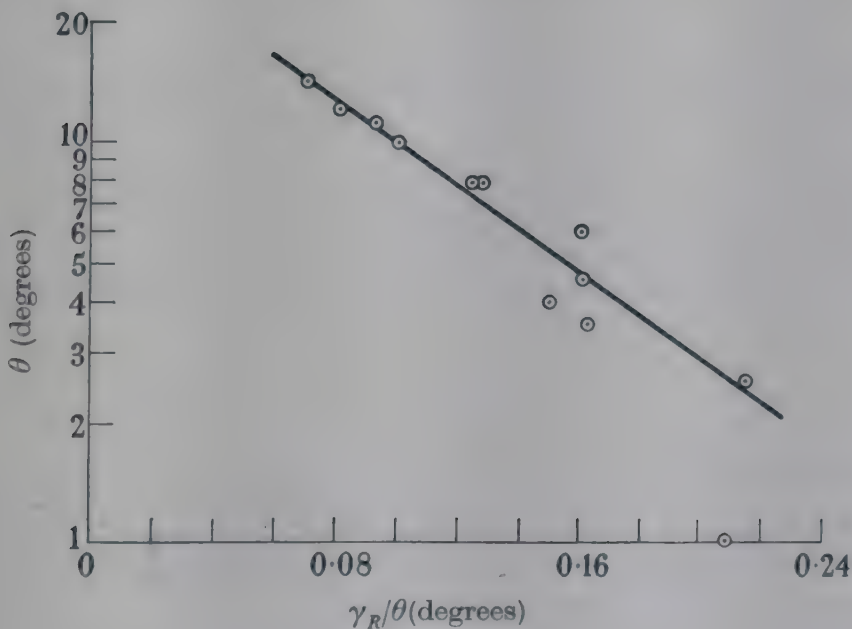


FIGURE 7.

If the dislocation theory can be applied to the grain boundary structure, then, as obtained from figure 6, the approximately constant maximum energies for lead and tin crystal boundaries are about 200 and 100 ergs per cm.² respectively. Shockley & Read estimate that Dunn & Lionetti's (1949) approximately constant value for silicon iron is 1300 ergs/cm.² using $A = +0.5$.

The authors wish to express their indebtedness to the National Research Council, Ottawa, and the School of Engineering Research of the University of Toronto, for financial assistance, and to the Consolidated Mining and Smelting Co., Trail, B.C., for the high-purity lead.

REFERENCES

- Aust, K. T. & Chalmers, B. 1950 *Proc. Roy. Soc. A*, **201**, 210.
 Boas, W. 1947 *Physics of metals*, p. 51.
 Bragg, W. L. 1940 *Proc. Phys. Soc.* **52**, 54.
 Burgers, J. M. 1940 *Proc. Phys. Soc.* **52**, 23.
 Chalmers, B. 1949 *Proc. Roy. Soc. A*, **196**, 64.
 Dunn, C. G. & Lionetti, F. 1949 *J. Metals Tech.* **1**, 125.
 Herring—unpublished (referred to by Dunn, Bolton & Daniels 1950
J. Metals Tech. **2**, 368).
 Shockley, W. 1949 *A.S.M., Cold working of metals*, p. 131.
 Shockley, W. & Read, W. T. 1949 *Phys. Rev.* **75**, sec. 4, 692.

The viscosity of liquid helium between 2 and 5° K

BY R. BOWERS AND K. MENDELSSOHN

Clarendon Laboratory, University of Oxford

(Communicated by F. E. Simon, F.R.S.—Received 19 July 1950)

The viscosity of liquid helium has been measured with a flow method between 2 and 5° K. The coefficient of viscosity decreases with decreasing temperature over the whole range. A particularly interesting feature is a sharp drop in the viscosity *above* the λ -point which coincides with an anomalous fall in the entropy in the same temperature interval. Possible explanations of this phenomenon have been discussed.

INTRODUCTION

The anomalous transport phenomena exhibited by liquid helium below 2.19° K (the so-called λ -point) are clearly manifestations of a new physical state into which the liquid passes at this temperature. The inception of this new state is accompanied by a rapid decrease in entropy. At present there does not exist a generally accepted theory of liquid helium, but it seems probable that this entropy decrease is due to an ordering process in momentum space. Well above the λ -point the entropy of the liquid changes roughly in proportion with the absolute temperature, while below it a high power of T (4–6) is followed. The shape of the specific heat curve (Keesom & Keesom 1935) shows, however, that the anomalous drop in the entropy begins at a considerably higher temperature ($\sim 2.6^\circ$ K) and that at 2.19° K about 8 % of the total entropy has already disappeared. The question therefore arises as to

whether this initial drop in entropy above the λ -point will show itself in the transport properties of liquid helium I below 2.6° K. If this were so, one may then ask whether the fall in entropy above 2.19° K is due to small regions having entropy values corresponding to helium II, or due to the appearance of a short-range ordering effect within the bulk of helium I.

The available data show little indication of any anomalous effects in this region. Pressure-independent flow (superfluidity) (Allen & Misener 1939), the thermo-mechanical effect (Allen & Reekie 1939) and the film transfer (Daunt & Mendelssohn 1939), all seem to approach the value zero as, on warming, the temperature 2.19° K is reached. On the other hand, the ordinary transport phenomena such as viscosity and thermal conductivity have, as yet, hardly been investigated in this temperature region.

The viscosity of liquid helium was first determined from observations on the torsional oscillations of a cylinder immersed in the liquid by Wilhelm, Misener & Clark (1935). The results obtained by this method do not agree even qualitatively with later determinations. They suggested a threefold rise in the coefficient of viscosity (η) with decreasing temperature from 4.22 to 2.19° K and then a rapid fall in η at the λ -point. Kapitza (1938) has suggested that these results correspond to turbulent motion due to the very small kinematical viscosity of liquid helium.

Later determinations by the oscillating disk method made in Leiden (Keesom & MacWood 1938; Keesom & Keesom 1941) indicate a steady decrease in η from 4.2 to 2.19° K followed by a very rapid decrease below the λ -point. The only previous determination of η by a flow method is that of Johns, Wilhelm & Grayson Smith (1939). These results show a rough qualitative similarity to the Leiden results, although the absolute values given by Johns *et al.* are much higher. However, only one point was measured (at 3.4° K) between the boiling-point and the λ -point.

It thus appeared that the viscosity of liquid helium was not known in sufficient detail in the temperature range from 2.19 to 2.6° K to allow comparison with the entropy curve. We have therefore investigated this region with a capillary flow method in order to determine not only the magnitude but also the character of the flow just above the λ -point.*

METHOD

The apparatus used consisted of a glass capillary of $\sim 4 \times 10^{-3}$ cm. radius and length 2.5 cm., one end of which was attached to a glass reservoir, while the other end was open to the helium bath (figure 1). The cross-section of the reservoir was a semicircle, so that it could be attached in a compact manner to a similar reservoir which was sealed at the lower end instead of being joined to a capillary. This second vessel was used to correct level readings for any variation due to evaporation.

After some liquid had been filled into both reservoirs, they were lifted out of the bath of liquid helium until the bath level was about 3 cm. below the level in the vessels. Observations were then made on the levels inside the two reservoirs and

* A short preliminary note on some of the results has already been published (Bowers & Mendelssohn 1949a).

on the outer level, using a cathetometer. The flow out of the capillary reservoir is governed by the relation

$$A \frac{dh}{dt} = \frac{\rho g H \pi a^4}{8l\eta},$$

$$\eta = \frac{\rho g \pi a^4}{8lA} \frac{H}{dh/dt} = \frac{K}{A} \rho g \frac{H}{dh/dt},$$

where a = radius of capillary,

l = length of capillary,

A = area of cross-section of reservoir,

dh/dt = rate of drop of level h in capillary reservoir,

H = difference between the levels inside the reservoir and outside it,

$K = \pi a^4/8l$ = constant for the capillary,

ρ = density of the liquid.

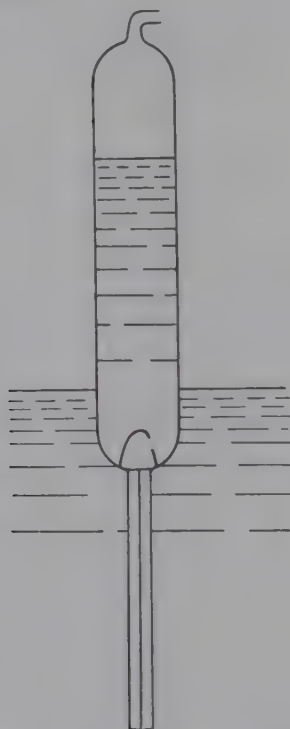


FIGURE 1. Capillary and reservoir.

From the curves relating h and t , dh/dt was calculated by drawing tangents. Generally six or seven values of dh/dt were calculated for pressure heads varying from about 3 to 1 cm. of helium for each determination of η . The values of dh/dt calculated in this way were corrected for the variation of h with t due to evaporation, by an additive correction calculated from observations on the second reservoir. This correction was of the order of 1 % for the highest rates of flow but increased to 6 % for the smallest flow rates. From each corrected value of dh/dt , individual values of $(dh/dt)/H$ were computed. For the six or seven determinations of this constant, the resulting deviation from the mean value had an average of about 3%. The Poiseuille constant K of the capillary was determined with flow of oxygen gas at room temperature. The values of ρ were taken from the determinations of Kamerlingh Onnes & Boks (1924) and Mathias, Crommelin, Kamerlingh Onnes & Swallow (1925). Since over the whole helium I region the flow was found to be independent of velocity, no corrections for end-effects or dissipation of kinetic energy were applied.

The experiments were carried out in a small Linde liquefier of the type described previously by Daunt & Mendelssohn (1948). The determinations of the level height were carried out by direct observation, the helium cryostat being surrounded by two Dewar vessels containing liquid hydrogen and liquid nitrogen respectively. The reservoirs were illuminated by a 12 W bulb placed behind a filter of copper sulphate solution in water at a distance of 1 m. from the cryostat. This illumination had no observable effect on the evaporation of helium. The total evaporation from the cryostat was of the order of 2 ml. of liquid per hour. The liquid helium I, which has a very poor thermal conductivity, was always stirred before a determination by moving the reservoirs up and down in order to accelerate the establishment of thermal equilibrium.

Three charcoal cleaners, two in liquid air and one in liquid hydrogen, were used to purify the helium gas before liquefaction. All the liquid helium used in the experiments had been obtained by the Joule-Kelvin expansion and was not derived from direct condensation of the gas. This ensured freedom from contamination by solidified particles of air (cf. Bowers & Mendelssohn 1949*b*). As an extra precaution, a small glass shield was placed over the top of the capillary and the top of the reservoir bent over (see figure 1) so that dust particles could not drop down into the vessel and fall straight into the capillary.

RESULTS

Measurements were carried out at temperatures between 2.12 and 5.0° K. Graphs of $\log dh/dt$ against $\log H$ were plotted for a number of points along the whole temperature range. At all temperatures above the λ -point, all curves indicated pure laminar flow. Immediately below the λ -point deviations from 'classical' flow began to occur. These results indicated a power law relating dh/dt and H with an index less than one for low pressure heads, followed by a fall in the value of this index at higher pressures, probably due to turbulence. These results are more in accord with those of Allen & Misener (1939) than with those of Johns *et al.* (1939), who tried to represent flow just below the λ -point as the sum of a Poiseuille flow and a pressure-independent flow. The values of η below the λ -point shown in figure 2 were calculated from the average values of $(dh/dt)/H$ observed at these temperatures. Hence these values represent an average viscosity over the pressure heads used and are only inserted for a comparison with values above the λ -point which had been measured over the same pressure heads.

The main results of this investigation, the temperature dependence of the viscosity in the helium I range, are given in figure 2. This shows that between 5 and 2.7° K, the viscosity falls only slightly from about $33\mu P$ ($1\mu P = 10^{-6}$ c.g.s.) to $30\mu P$. Then follows a sharp drop down to a value of $22\mu P$ at 2.19° K. Although below the λ -point the flow becomes increasingly non-classical, our values near 2.19° K do not indicate discontinuity in η at the λ -point. The question of whether a discontinuous change in the slope of the η - T curve occurs at the λ -point cannot be decided with certainty by our experiment, but it can be seen (cf. figure 3) that any change in slope compatible with our results would not be nearly as pronounced as that found in Keesom's

oscillation experiments. The scatter of our values, particularly at the higher temperatures, is slightly larger than can be expected from the accuracy of observation and computation. This is possibly due to the passage of minute gas bubbles through the capillary. These bubbles were never large enough to be visible, but their presence in the liquid is indicated by the occasional formation of visible bubbles under immersed horizontal surfaces.

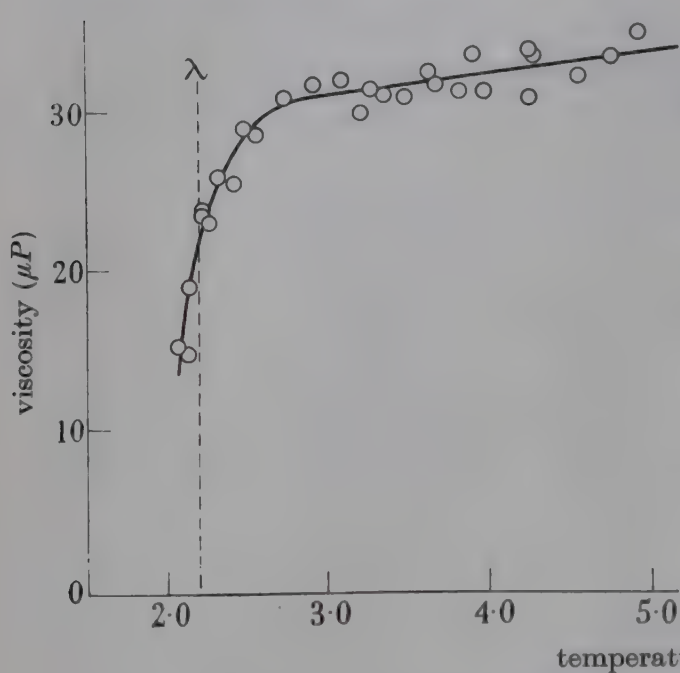


FIGURE 2. Viscosity of liquid helium.

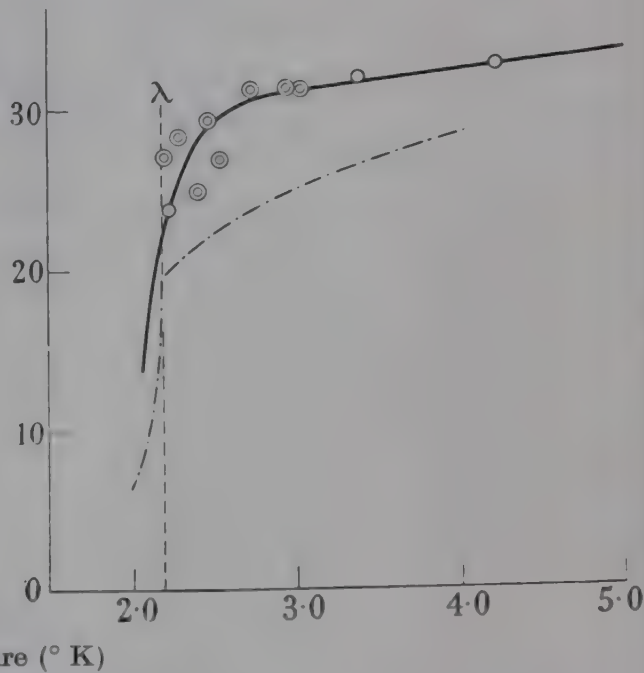


FIGURE 3. Comparison of viscosity measurements. — present work (glass capillary). $\odot$ present work (wire tube). $\circ$ Johns *et al.* (multiplied by 0.7). --- Keesom & Keesom, Keesom & MacWood (oscillating disk).

The absolute values given for η by us are more in accord with the Leiden measurements than with the measurements of Johns *et al.* In the latter paper no independent calibration experiment is mentioned, and the absolute values may have been computed from the measured dimensions of the capillary which can easily account for the discrepancy since η depends on the fourth power of the radius. There is, however, an important qualitative difference between the Leiden results and ours, in that the former gave no indication of the steep fall in η in the same region (just above the λ -point) where the entropy of the liquid begins to decrease rapidly. The three significant points given by Johns *et al.* for the helium I region do fit our curve if the absolute values are adjusted to agree for one point. The results of these three investigations are compared in figure 3.

One may be inclined to ascribe this fall in our values of η just above the λ -point to the first appearance of superfluidity. However, as mentioned above, observations of dh/dt in this region gave no indication of non-classical flow. Since the effect of a small degree of superfluidity might be swamped in viscosity measurements in a wide capillary, a more sensitive method was used in order to investigate this point.

The glass capillary used in the previous experiment was replaced by a cupro-

nickel tube containing 900 wires of 46 s.w.g. Eureka. This metal tube, of the same length as the capillary, had been pulled through a succession of dies until its Poiseuille constant was approximately equal to that of the glass capillary. Thus the wire tube would offer approximately the same resistance to the bulk flow of helium I as the glass capillary, but due to the much smaller size of the individual channels (~ 10 to 100 times) it would give greatly increased flow once superfluidity came into play.

The tube was joined by a small copper-glass seal to the same reservoir as was used in the capillary experiment. A sealed off reservoir was also employed to determine the evaporation correction as described in the capillary experiment. The initial observations on the flow through this tube established the fact that it was difficult to get helium I to flow through the tube at temperatures above $\sim 3^\circ\text{K}$ unless the liquid was first cooled below 3°K and then allowed to warm to the higher temperature. Initially, we associated this difficulty with the large number of gas bubbles collecting at the lower end of the tube. In order to reduce this a shield was fitted below the tube to deflect rising bubbles from the end of the tube. In spite of this precaution, the tube often blocked above 3°K , although less difficulty was met below 3°K . These blockages were probably caused by the presence of small gas bubbles in the wire channels. Compared with the glass capillary, surface-tension forces are much increased in using this wire tube due to a decrease in the effective radius of the flow channels. The fact that the blockages disappear below 3°K may indicate a rise in the thermal conductivity of the helium in this region. Only a single value of the heat conductivity of helium I (at 3.3°K) has so far been measured (Keesom & Keesom 1936), and it is quite possible that the drop in viscosity below 2.6°K is accompanied by a fall in thermal resistance. Determinations of the heat conductivity in this region are now in progress.

Owing to the difficulties experienced with narrow channels in helium I, the points show a greater scatter than those obtained with the wide capillary, but they are sufficiently accurate to allow a decision on the question of superfluidity. The values obtained for η using this tube are shown in figure 3 superimposed on the η - T curve determined with the glass capillary. The wire tube was calibrated using the results of the previous experiment by observing the flow of liquid helium through it at 2.99°K and assuming that η was equal to $31.2\mu\text{P}$ at this temperature. These results show that no greatly increased flow seems to occur as the temperature is reduced from 3°K to the λ -point; the points obtained with the wire tube do not indicate a more rapid fall than those obtained with the capillary. On the other hand, when carrying out the experiment just below the λ -point at 2.12°K , the flow through the tube was increased to at least 6 times that at 2.26°K , and because of its rapidity accurate observation of the flow became impossible. Thus the wire tube does act as a sensitive indicator of superfluidity although it is not so reliable for an overall determination of η .

DISCUSSION

The results obtained in this work show that the appearance of the anomalous liquid helium II is preceded by a rapid drop in the viscosity of helium I within

$\frac{1}{2}$ degree above the λ -point. This temperature region is the same as that in which the entropy shows the first anomalous decrease. However, the comparison experiment with flow in narrow channels shows clearly that the drop in viscosity cannot simply be attributed to the inception of superfluidity. All experiments on the latter phenomenon, including the values observed in this investigation below the λ -point, emphasize the sensitivity of superfluid transport to the size of the flow channel. Moreover, the lack of a non-classical term in all our observations above the λ -point emphasizes the fact that the anomalous high-flow rates in helium I are not due to the appearance of frictionless flow. In contradistinction to the transport phenomena in liquid helium II, our results suggest that above the λ -point we meet an anomalous drop in the true viscosity. Even so, the discrepancy between our results obtained with a flow method and those measured with an oscillating disk must appear disturbing. In the latter experiments, the sharp drop in η below 2.6°K does not seem to occur, and instead there is a rapid change, even if not a discontinuity, at the λ -point. Since the three determinations of η by a flow method obtained in Toronto agree with our curve (except for the absolute value), the impression is gained that even the true viscosity of helium I may depend on the method of measurement. Only further determinations by an oscillation method could settle this particular question.* The negative result with a wire tube also seems to indicate that the drop in the viscosity of helium I is not due to thermal fluctuations, causing the appearance of small regions of helium II within the liquid. A further argument against such a concept of fluctuations is provided by the relatively large range of absolute temperatures above the λ -point which exhibit the anomalous fall of the viscosity. Using the standard fluctuation formula $\overline{\Delta T^2} = kT^2/C$, where C is the specific heat of the elementary volume considered, it appears that helium II regions of the linear dimensions of the film thickness (~ 100 atoms) would not be noticeable for more than 10^{-3} degree above the λ -point. In order to produce an effect of the order $\Delta T = 0.4$ degree the regions would have to be as small as 5 atoms diameter.†

A more probable explanation for the drop in η suggests itself when we consider that the fall in viscosity in helium I is accompanied by a drop in entropy, i.e. by an increase in the degree of order of the liquid. Since X-ray photographs (Keesom & Taconis 1938) and an optical investigation (Schubnikow & Kikoin 1936) have completely failed to give evidence of a space order in helium II, it is generally assumed that the drop in entropy is indicative of an ordering process in momentum space. The recent researches into the properties of liquid helium with the atomic weight 3 lend support to London's (1938) hypothesis that the λ -phenomenon corresponds to the condensation of a Bose-Einstein gas. It has proved so far impossible to extend the theoretical considerations rigorously from an ideal gas to a liquid with appreciable

* *Note added in proof.* We have been informed that such experiments have now been carried out in Leiden (Smith 1950) and that the new results obtained with the oscillating disk method are in qualitative agreement with the present work.

† It is possible, though rather unlikely, that the specific heat of the liquid would be dependent on a subdivision into small regions or that the viscosity may depend on another parameter (e.g. the pressure) which might be very sensitive to fluctuations. In both cases our rough assessment of fluctuations would become invalid. We are grateful to Dr D. K. C. MacDonald for discussions on this problem.

interaction. However, accepting the Bose-Einstein model as basically correct, one must correlate the entropy drop above the λ -point with a similar drop in the entropy of the ideal gas preceding the actual condensation phenomenon. This increase in the state of order of the ideal gas, which extends over a considerable range of temperature, corresponds to the change from a classical distribution of velocities to that of a degenerate Bose-Einstein gas. The latter is characterized by a heavy occupation of the lowest energy states, which, in the case of the liquid, are probably separated from the ground state by an energy gap. Consequently one is inclined to ascribe the sharp drop in viscosity above the λ -point to a redistribution in the velocity spectrum of the liquid, preceding the actual condensation process. A parallel effect occurs in the order-disorder transformation in solids where the similar decreases in entropy above the λ -point are interpreted as the establishment of short range order preparatory to long-range order at lower temperatures (cf. Fowler 1936).

The shape of the viscosity curve suggests two separate phenomena; one responsible for the rapid drop in η below 2.6° K while the other causes a slow decrease over the whole temperature range. It is impossible to say to what degree these phenomena are connected. In an ordinary liquid one would expect the viscosity to rise rather than to fall with decreasing temperature. However, quite apart from the λ -phenomenon liquid helium is not an ordinary liquid since its density is four times smaller than would be expected from the gas-kinetic diameter of the helium atom. Simon (1934) has pointed out that the high atomic volume is due to the high zero-point energy which incidentally prevents the formation of crystalline helium under its own vapour pressure. Thus even in liquid helium λ quantum effects are predominant, and it is possible that the gradual fall in viscosity between 5 and 3° K is directly caused by the high zero-point energy. Whether this is the case or whether this gradual drop, too, has to be considered as a preparation for the λ -phenomenon could be decided by comparison with the lighter isotope. Liquid ^3He has a zero-point energy similar to that of ^4He , but it does not show the λ -phenomenon. Unfortunately, published data on the viscosity of liquid ^3He (Osborne, Weinstock & Abraham 1949) are not sufficient to allow a clear decision as to whether in this case η rises or falls with decreasing temperature.

One of us (R.B.) is indebted to the D.S.I.R. for a grant enabling him to carry out this research.

REFERENCES

- Allen, J. F. & Misener, A. D. 1939 *Proc. Roy. Soc. A*, **172**, 467.
 Allen, J. F. & Reekie, J. 1939 *Proc. Camb. Phil. Soc.* **35**, 114.
 Bowers, R. & Mendelssohn, K. 1949a *Proc. Phys. Soc. A*, **62**, 394.
 Bowers, R. & Mendelssohn, K. 1949b *Nature*, **163**, 870.
 Daunt, J. G. & Mendelssohn, K. 1939 *Proc. Roy. Soc. A*, **170**, 423.
 Daunt, J. G. & Mendelssohn, K. 1948 *J. Sci. Instrum.* **25**, 318.
 Fowler, R. H. 1936 *Statistical mechanics*, 2nd ed. Cambridge University Press.
 Johns, H. R., Wilhelm, J. O. & Grayson Smith, H. 1939 *Canad. J. Res. A*, **17**, 149.
 Kamerlingh Onnes, H. & Boks, J. D. A. 1924 *Commun. Phys. Lab. Univ. Leiden*, no. 170b.
 Kapitza, P. 1938 *Nature*, **141**, 74.
 Keesom, W. H. & Keesom, A. P. 1935 *Physica, 's Grav.*, **2**, 557.

- Keesom, W. H. & Keesom, A. P. 1936 *Physica, 's Grav.*, **3**, 359.
 Keesom, W. H. & Keesom, P. H. 1941 *Physica, 's Grav.*, **8**, 65.
 Keesom, W. H. & MacWood, G. E. 1938 *Physica, 's Grav.*, **5**, 737.
 Keesom, W. H. & Taconis, K. W. 1938 *Physica, 's Grav.*, **5**, 270.
 London, F. 1938 *Nature*, **141**, 643.
 Mathias, E., Crommelin, C. A., Kammerlingh Onnes, H. & Swallow, J. C. 1925 *Commun. Phys. Lab. Univ. Leiden*, no. 172b.
 Osborne, D. W., Weinstock, B. & Abraham, B. M. 1949 *Phys. Rev.* **75**, 888.
 Schubnikow, L. W. & Kikoin, A. K. 1936 *Phys. Z. U.S.S.R.* **10**, 119.
 Simon, F. 1934 *Nature*, **133**, 529.
 Smith, P. L. 1950 *Physica, 's Grav.*, (in the Press).
 Wilhelm, J. O., Misener, A. D. & Clark, A. R. 1935 *Proc. Roy. Soc. A*, **151**, 342.

The calculation of deflexions of the vertical from gravity anomalies

By A. H. COOK

Department of Geodesy and Geophysics, University of Cambridge

(Communicated by R. Stoneley, F.R.S.—Received 20 July 1950—Revised 30 August 1950)

The theory of the application of gravity measurements to geodetic calculations is discussed, and the errors involved in calculating deflexions of the vertical are estimated. If the gravity data are given as free air anomalies from Jeffreys's (1948) formula, so that the second and third harmonics of gravity are assumed known, the orders of magnitude of the standard deviations of the different sources of error are the following:

Single deflexion: neglect of gravity outside 20°	1"
Difference of deflexions: neglect of gravity outside 5°	0".5
Calculation of effects of gravity from 0.05° to 5°	0".1
Calculation of effects of gravity within 0.05°	between 0".1 and 0".5

Estimates of the deflexions are made for Greenwich, Herstmonceux, Southampton and Bayeux, and the difference between Greenwich and Southampton is compared with the astronomical and geodetic amplitudes.

1. INTRODUCTION

It is more than one hundred years since Sir George Stokes (1849) showed how the form of the sea-level surface could be calculated from knowledge of gravity over the surface of the earth, yet despite its fundamental importance for geodesy very few applications of his theorem have been made (Meinesz 1928; Steenwijk 1948; Rice 1949; Lambert & Darling 1936). The reason is simple: the existing observations of gravity are few and unsatisfactorily distributed (Hunter 1935). Until recently this was true even of the British Isles, for although they were, by comparison with the rest of the world, well surveyed, yet they contained only the pendulum stations described by Bullard & Jolly (1936) and are surrounded by seas, in which gravity measurements are of necessity made at rather large intervals. But within the last ten years many observations have been made in Great Britain

(White 1949; Cook & Thirlaway 1950) and Ireland (Cook, Murphy & Thirlaway 1950), in the Atlantic Ocean and English Channel (Browne & Cooper 1950), and in northern France (Lejay 1947), and have made it worth while to study the application of Stokes's theorem in this country. The object of this paper is therefore to examine the accuracy attainable in the calculations with the existing distribution of data in and around the British Isles. The method adopted is to calculate the difference of deflexions of the vertical at pairs of stations and to compare it with the difference of amplitude as determined astronomically and geodetically. For the comparison to have any significance, the expected uncertainty of the calculation must be estimated, and a discussion of the sources of error therefore forms a considerable part of this paper. These errors depend to some extent on the manner in which the gravity observations are reduced, for if the reduction is incorrect systematic errors may be introduced. The principles of the calculations are therefore considered, and the reasons for adopting the method of reduction that has been used, the free-air reduction, are explained.

A deflexion of the vertical at any point is the angle between the normal to the surface of constant geopotential through the point and the normal to a conventional surface, such as the International Spheroid, to which geodetic measurements are reduced. It has components ξ in the meridian and η in the prime vertical, and from Stokes's expression for the relation of gravity to the form of the geoid (Stokes 1849) it may be shown that

$$\frac{\xi}{\eta} = -\frac{1}{2\pi g} \int_0^\pi \int_0^{2\pi} \Delta g \frac{\partial f}{\partial \psi} \sin \psi \frac{\cos \chi}{\sin \chi} d\psi d\chi$$

(Meinesz 1928; Hunter 1935).

Δg is the free-air anomaly of gravity at the point distant ψ , azimuth χ .

This expression involves neglect of terms of order $(h/a)^2$, where h is the height of the station above sea-level and a is the radius of the earth (Jeffreys 1931, 1932, 1933). One of these terms, discussed in the appendix, arises from the calculation of the deflexions in effect not at the external point but on the geoid.

2. THE DATA TO BE USED IN STOKES'S FORMULA

Although the geodetic problem is formally soluble by Stokes's method, there are difficulties in the practical application which arise from the scarcity of gravity observations in large tracts of the earth's surface, and from the fact that observations are rarely equally numerous on high and low ground.

The computation of ξ and η can conveniently be performed by constructing templates on which radii and concentric circles describe compartments which give equal contributions to the double integrals for equal mean values of Δg in the compartment (see § 5, and Sollins 1947), and the interpolation difficulty arises through the formation of these means. Ways of avoiding the difficulty will now be considered.

A general method is the following. Let the free air anomaly Δg be broken down into two parts, Δg_r and Δg_c , where Δg_c is some calculable term, depending on the

topography, and Δg_r is the residual which might be, for example, the Bouguer anomaly or an isostatic anomaly. Then obviously

$$\iint \Delta g \frac{\partial f}{\partial \psi} \frac{\cos \chi}{\sin \chi} dS = \iint \Delta g_r \frac{\partial f}{\partial \psi} \frac{\cos \chi}{\sin \chi} dS + \iint \Delta g_c \frac{\partial f}{\partial \psi} \frac{\cos \chi}{\sin \chi} dS,$$

which means that instead of working directly with the free-air anomalies, the two components, the calculated and the residual parts, may be integrated separately. Thus, if in a given area the Bouguer anomalies do not vary very much, they could be used for Δg_r , and the attraction of the topography, which has to be added to $\iint \Delta g_r \frac{\partial f}{\partial \psi} \frac{\cos \chi}{\sin \chi} dS$, could be found by the formulae given by Clarke (Clarke 1858, see § 5). The Bouguer anomalies are only useful in rather small areas, for over large regions they are often more variable than the free-air anomalies. For this reason the use of isostatic anomalies has been proposed (see, for example, Tanni 1948). This proposal has given rise to a certain amount of confusion because of attempts to produce a physical interpretation of the term

$$\iint \Delta g_r \frac{\partial f}{\partial \psi} \frac{\cos \chi}{\sin \chi} dS.$$

It is by no means clear what this interpretation should be, for, among other difficulties, Jeffreys has pointed out (1948) that the application of the isostatic corrections alters the centre of mass of the earth and so introduces a forbidden first-order harmonic term into gravity. However, deflexions on a hypothetical compensated planet are not desired, but those on the actual earth, and the physical interpretation, supposing there to be one, of the two parts into which the free-air anomaly is decomposed, is irrelevant provided the parts are ultimately recombined to give the deflexions of the actual external equipotential surface. There is thus no theoretical difficulty in the correct use of the isostatic anomalies and the real question is whether there is any practical advantage.

The term

$$\iint \Delta g_c \frac{\partial f}{\partial \psi} \frac{\cos \chi}{\sin \chi} dS$$

is the deflexion due to the topography and compensation and may be calculated with the aid of the Fundamental Tables calculated by the U.S. Coast and Geodetic Survey (Darling 1948). It remains to decide whether the variance of isostatic anomalies is so much less than that of the free-air anomalies that the extra facility of performing Stokes's integration compensates for the additional calculation of the deflexion of the topography and compensation, and whether it is easier to make some assumption for isostatic anomalies in unsurveyed regions than for the free-air anomalies.

The data used in investigating these questions are the isostatic anomalies (Hayford, $D = 113.7$ km.) assembled by Tanni (1948, maps II and III). From them the variances of the means of 10° and 30° squares (the analogues of Jeffreys's τ_1 and τ_2 variances of the free-air anomalies (Jeffreys 1941, 1943)) have been calculated. Tanni gives the means over 1° squares for Europe, North Africa and Asia, and

the means over 5° squares for the whole world. From the former the standard deviation of a 1° square within a 10° square is found to be approximately 12 mgals, and the uncertainty of the mean of a 10° square from this cause is therefore 1.2 mgals. From the 5° data, τ_1 , the variance of the means of 10° squares within the 30° squares is found to be 184 mgals². The effect of the standard deviation of the 1° squares has been neglected, as is obviously permissible. The estimate of τ_2 , the variance between 30° squares, is 110 mgals². These results are not exactly comparable with Jeffreys's statistics, for Tanni did not select his data in the same way as Jeffreys; in forming the means over 10° squares the effect of height has been neglected, and the 30° squares are those between Lat. 60° N. and 60° S.

The variance of the free-air anomalies between 1° squares leads to the mean values of 10° squares having standard deviations whose median is 14 mgals; with isostatic anomalies, the standard deviation of the means of 10° squares due to the variance between 1° squares is 1.2 mgals, which is very much less. Accordingly, within small distances from the point at which deflexions are to be calculated, it would be very much easier to interpolate the isostatic anomalies instead of the free-air anomalies. But it is the scarcity of measurements at large distances which gives the greater difficulty, and for these there is little advantage in the use of isostatic anomalies, for the τ_1 variance is 184 against 433 mgals² for the free-air anomalies, and the τ_2 variance is 110 against 227 mgals². The difference between the latter cannot be considered significant, since both the free-air and isostatic variances are based on only about 30 degrees of freedom. These results are not the most favourable to the isostatic hypothesis, since the Hayford scheme, $D = 113.7$ km., is certainly not the most successful in many regions, but rearrangements of masses within the first 100 km. of the earth's crust will not affect the means of 10° and 30° squares at all significantly, and will not have much effect even within 5° squares. Thus the present comparisons are sufficient to determine whether there is any practical advantage in using isostatic anomalies and evidently there is not. A further consideration is that the mean free-air anomaly over the whole world is forced to be zero, so that this is the best value to assume in an unsurveyed region. The mean value of the isostatic anomalies given by Tanni's data is +4 mgals, so that it is not obvious that isostatic anomalies should be assumed zero in the absence of observations. Thus the conclusion is that while there is no theoretical objection to the use of isostatic anomalies, there is no practical advantage, either in a much greater facility of interpolation, or in making assumptions about anomalies in unsurveyed regions.

Since the use of isostatic anomalies involves the labour of calculating them, in addition to that of computing the deflexion due to topography and compensation, it is certainly not worth while to use them for any local area in which the Bouguer anomalies behave smoothly, for if the Bouguer anomalies are used, the only additional calculation is that of the deflexion due to the topography. Thus, if the direct use of free-air anomalies is not satisfactory, Bouguer anomalies may be advantageous. Another way of using Bouguer anomalies which is useful if the topography is not too rugged is to construct a map of free-air anomalies by combining maps of Bouguer anomalies and topography; this has been done for the Wealden district (see § 5).

3. THE ERRORS INVOLVED IN CALCULATING ABSOLUTE DEFLEXIONS

3.1. *The use of Jeffreys's determination of the earth's gravitational field and the restriction of Stokes's integral to a local cap*

Since except in fairly small regions none of the expedients discussed above appear to have any advantage over the direct use of the free-air anomalies, the most efficient means of interpolating these will now be considered. Clearly this will be a spherical harmonic development of the anomalies with the coefficients determined by least squares, and this is just the problem that Jeffreys has solved (1941, 1943). But for the harmonics which can be so determined, there is no point in using Stokes's formula because given the departure of gravity from the International formula, for which Jeffreys's solution is

$$\Delta g_0 = +2.5 - 6.1 P_2 + 4.0 P_2^2 \cos 2\lambda + 1.3 P_3^2 \cos 2\lambda + 4.2 P_3^1 \cos \lambda + 0.46 P_3^3 \sin 3\lambda,$$

the departure of the corresponding reference surface from the International Spheroid is also known; in metres it is

$$26.1 P_2^2 \cos 2\lambda + 4.2 P_3^2 \cos 2\lambda + 13.8 P_3^1 \cos \lambda + 1.50 P_3^3 \sin 3\lambda$$

(the P_n^m are functions of latitude: λ is longitude from Greenwich).

The contributions from these harmonics to the deflexions ξ , η , in the meridian and prime vertical, $\partial N/a \partial \theta$ and $\partial N/a \cos \theta \partial \lambda$, are

$$\begin{aligned} \Delta \xi &= \frac{10^{-6}}{6.4} [52.2 \cot \theta P_2^2 \cos 2\lambda + 13.8 \cot \theta \cos \lambda P_3^1 \\ &\quad + (8.4 \cot \theta \cos 2\lambda - 13.8 \cos \lambda) P_3^2 + (4.5 \cot \theta \sin 3\lambda - 4.2 \cos 2\lambda) P_3^3], \\ \Delta \eta &= \frac{10^{-6}}{6.4 \cos \theta} [-52.2 P_2^2 \sin 2\lambda - 8.4 P_3^2 \sin 2\lambda - 13.8 P_3^1 \sin \lambda + 4.5 P_3^3 \cos 3\lambda]. \end{aligned}$$

If the free-air anomalies are now given as residuals from Jeffreys's gravity formula, the results of Stokes's integration applied to these anomalies will be deviations from Jeffreys's reference surface. Apparently there is no practical advantage in the use of such residual anomalies, for Stokes's integration has still to be performed, but the variance of the residual anomalies will be less than that of the anomalies from the International formula, so that interpolation should be easier. More important, since the effect of anomalies at distant points is mainly due to the lower harmonics, which Jeffreys has determined, it may be possible to restrict the range of Stokes's integral to a fairly limited cap centred on the point at which the deflexion is required. At present the advantages are not very great, for of the total variance of the means of 10° squares, 660 mgals², only 98 mgals² is the contribution of the harmonics found by Jeffreys, so that it is no easier to interpolate the residuals than the anomalies from the International formula. The obstacle to the determination of more harmonics is the lack of observations in the Pacific Ocean and southern hemisphere. Also, as will be shown, it is at present necessary to extend Stokes's integral over about 20° radius, but with higher-order harmonics determined it could probably be restricted to 10° or less.

The error committed in limiting Stokes's integral to a given cap is therefore now studied. The objects are to estimate the errors which would be committed with the

present gravity data, assumed to be defined by the harmonics fitted by Jeffreys together with the τ_1 and τ_2 variances, and to examine the effect of being able to fit higher harmonics, the probable consequence of suitably distributed new observations. These are different objects from those of Hunter (1935), who examined the errors likely with an ideal distribution of stations and with integrals taken over the whole world.

The error committed in limiting Stokes's integral to a given cap

Assume the free-air anomaly from Jeffreys's formula, Δg , to be expanded in spherical harmonics about the point at which a deflexion is to be found, that is,

$$\Delta g = \sum_{n=4}^{\infty} \sum_{m=0}^n (a_{nm} \cos m\chi + b_{nm} \sin m\chi) P_n^m(\cos \psi).^*$$

The series begins at $n = 4$ because Jeffreys has found the lower harmonics, which are therefore excluded from the discussion.

Since the coefficients N_{nm} in the spherical harmonic expansion of the departure N of the geoid from the reference surface are

$$-(a_{nm}, b_{nm}) \frac{a}{(n-1)g},$$

the deflexion due to a single harmonic is found by differentiating N in the appropriate direction and putting ψ equal to 0 to be

$$\frac{\xi_n}{\eta_n} = \frac{1}{2g} \frac{n(n+1)}{n-1} \frac{a_{n1}}{b_{n1}}.$$

But it is the value of ξ_n (or η_n) when the integration with respect to ψ is carried not from 0 to π but from 0 to α that is required. This is called $\xi_n(\alpha)$, and is

$$-\frac{a_{n1}}{2g} \int_0^\alpha P_n^1 \frac{\partial f}{\partial \psi} \sin \psi d\psi.$$

The integral cannot be evaluated exactly but a numerical summation has been made.

Write the error $(\xi_n - \xi_n(\alpha))$ as $\epsilon_n(\alpha)$.

Then the sum of the squares $\sum_{n=4}^{\infty} \epsilon_n^2(\alpha)^2$ is required, and to evaluate it some assumption about the distribution of Δg among the different harmonics has to be made, since that distribution is quite unknown.

The assumption is, that each harmonic of degree n ,

$$\sum_{m=0}^n (a_{nm} \cos m\chi + b_{nm} \sin m\chi) P_n^m(\cos \psi),$$

* The definition used is

$$P_n^m(z) = \frac{(1-z^2)^{\frac{1}{2}m}}{2^n \cdot n!} \frac{d^{n+m}}{dz^{n+m}} (z^2-1)^n.$$

TABLE 1. $\int_{\alpha}^{\pi} P_n^1 \frac{\partial f}{\partial \psi} \sin \psi \, d\psi.$

α n	0°	2°	4°	6°	8°	10°	12°	14°	16°	18°	20°	22°	24°	26°	28°	30°
4	- 6.67	- 6.31	- 5.93	- 5.54	- 5.13	- 4.71	- 4.28	- 3.84	- 3.39	- 2.95	- 2.51	- 2.08	- 1.67	- 1.28	- 0.92	- 0.58
5	- 7.50	- 6.96	- 6.40	- 5.82	- 5.22	- 4.60	- 3.98	- 3.37	- 2.76	- 2.17	- 1.61	- 1.09	- 0.60	- 0.17	+ 0.20	+ 0.52
6	- 8.40	- 7.65	- 6.86	- 6.06	- 5.24	- 4.41	- 3.59	- 2.80	- 2.04	- 1.34	- 0.69	- 0.13	+ 0.36	+ 0.75	+ 1.04	+ 1.25
7	- 9.33	- 8.33	- 7.29	- 6.23	- 5.16	- 4.11	- 3.10	- 2.14	- 1.26	- 0.49	+ 0.17	+ 0.71	+ 1.11	+ 1.37	+ 1.51	+ 1.53
8	- 10.29	- 9.00	- 7.67	- 6.33	- 5.00	- 3.72	- 2.51	- 1.42	- 0.48	+ 0.31	+ 0.91	+ 1.33	+ 1.57	+ 1.64	+ 1.56	+ 1.36
9	- 11.25	- 9.64	- 7.99	- 6.34	- 4.74	- 3.23	- 1.87	- 0.69	+ 0.27	+ 0.99	+ 1.47	+ 1.70	+ 1.72	+ 1.55	+ 1.25	+ 0.86
10	- 12.22	- 10.26	- 8.26	- 6.28	- 4.40	- 2.68	- 1.19	+ 0.03	+ 0.94	+ 1.52	+ 1.79	+ 1.79	+ 1.55	+ 1.16	+ 0.68	+ 0.18
11	- 13.20	- 10.85	- 8.46	- 6.14	- 3.98	- 2.08	- 0.50	+ 0.69	+ 1.47	+ 1.85	+ 1.87	+ 1.61	+ 1.14	+ 0.57	0.00	- 0.48
12	- 14.18	- 11.41	- 8.62	- 5.94	- 3.51	- 1.44	+ 0.16	+ 1.26	+ 1.86	+ 1.98	+ 1.72	+ 1.21	+ 0.56	- 0.08	- 0.62	- 0.96
13	- 15.17	- 11.94	- 8.71	- 5.67	- 2.98	- 0.80	+ 0.78	+ 1.72	+ 2.07	+ 1.90	+ 1.38	+ 0.66	- 0.07	- 0.67	- 1.04	- 1.15
14	- 16.15	- 12.43	- 8.74	- 5.33	- 2.42	- 0.17	+ 1.32	+ 2.05	+ 2.10	+ 1.64	+ 0.89	+ 0.06	- 0.64	- 1.08	- 1.20	- 1.02
15	- 17.14	- 12.89	- 8.73	- 4.95	- 1.83	+ 0.44	+ 1.78	+ 2.23	+ 1.97	+ 1.24	+ 0.32	- 0.51	- 1.07	- 1.25	- 1.07	- 0.63
16	- 18.13	- 13.32	- 8.66	- 4.52	- 1.23	+ 1.00	+ 2.13	+ 2.27	+ 1.70	+ 0.74	- 0.25	- 0.97	- 1.28	- 1.16	- 0.70	- 0.10
17	- 19.12	- 13.72	- 8.54	- 4.05	- 0.63	+ 1.50	+ 2.37	+ 2.17	+ 1.30	+ 0.19	- 0.76	- 1.27	- 1.27	- 0.84	- 0.19	+ 0.43
18	- 20.12	- 14.09	- 8.38	- 3.54	- 0.04	+ 1.94	+ 2.49	+ 1.95	+ 0.82	- 0.34	- 1.15	- 1.37	- 1.04	- 0.37	+ 0.34	+ 0.81

$\partial f/\partial \psi$ is the derivative of Stokes's function tabulated by Sollins (1947).

$P_n^1(\cos \psi)$ is the associated Legendre function of the first kind.

The table is correct to within one unit of the last figure.

For interpolation Δ''' is negligible both n -wise and ψ -wise.

The table was computed by Miss C. Munford of the University Mathematical Laboratory, Cambridge; a manuscript table of P_n^1 for n up to 18 was also compiled.

makes an equal contribution to the variance of the anomaly, and also that each component of the n th degree harmonic,

$$a_{nm} \cos m\chi P_n^m \quad \text{or} \quad b_{nm} \sin m\chi P_n^m,$$

contributes equally to the variance of the n th degree harmonic.

The total variance used is the sum of the τ_2 and τ_1 variances, for the variations which do not affect the mean of a 10° square can be neglected and the effects of harmonics of degree 4 to 18 will be considered, that is, all those which keep the same sign over a 10° square, and which Jeffreys could not determine. The sum of the τ_1 and τ_2 variances is 570 mgals^2 . (The variance of the harmonics determined by Jeffreys has been subtracted.)

The number of harmonics is 15, so that the variance contributed by each is

$$\frac{570}{15} = 38 \text{ mgals}^2.$$

Hence each component of degree n contributes $38/(2n+1) \text{ mgals}^2$ to the total variance. But the variance due to the term

$$a_{n1} \cos \chi P_n^1 (\cos \psi)$$

is
$$\frac{a_{n1}^2}{4\pi} \int_0^{2\pi} \int_0^\pi \cos^2 \chi \{P_n^1\}^2 \sin \psi d\psi d\chi = \frac{n(n+1)}{2(2n+1)} a_{n1}^2.$$

Hence
$$a_{n1}^2 = \frac{76}{n(n+1)} \text{ mgals}^2.$$

The total deflexion due to the n th harmonic is

$$\xi_n = \frac{8.72 \sqrt{\{n(n+1)\}}}{2g(n-1)},$$

and the total deflexion, $(\sum \xi_n^2)^{\frac{1}{2}}$, is about $4''$ if the summation is for n from 4 to 18.*

We can now calculate $\epsilon_n(\alpha)$ and make an estimate of $\sum \epsilon_n^2(\alpha)$. The elements of the working are set out so that if it is desired to make some other assumption about the distribution of the coefficients of the harmonics, a_{n1} , the data will be available for the necessary alterations in the calculations. In table 1 the values of

$$\int_\alpha^\pi P_n^1 \frac{\partial f}{\partial \psi} \sin \psi d\psi,$$

$$\frac{n(n+1)}{n-1} - \int_0^\alpha P_n^1 \frac{\partial f}{\partial \psi} \sin \psi d\psi,$$

that is,

are given for values of n between 4 and 18 and values of α at 2° intervals between 0° and 30° .

The value of $\epsilon_n(\alpha)$ is the corresponding entry in table 1 multiplied by $a_{n1}/2g$, where it is assumed that $a_{n1} = 8.72/\sqrt{\{n(n+1)\}}$. The sums, $\sum \epsilon_n^2(\alpha)$, that result, and the corresponding uncertainties of the deflexion, are given in table 2.

TABLE 2

α	10°	15°	20°	25°	30°
$\sum \epsilon_n^2(\alpha)$	3.03	1.30	0.54	0.27	0.16
$\sigma(\xi)$	1.6	1.1	0.7	0.5	0.4

* The value estimated by Jeffreys (1948) for the root mean square deflexion for all harmonics is $3.8''$.

Now when these values of $\sigma(\xi)$ are compared with the r.m.s. value of the total deflexion, about 4", it is clear that even with the present unsatisfactory distribution of observations, quite useful results can be obtained with Stokes's integral restricted to 20°. If fourth and fifth harmonics could be determined, there would be two consequences: first, the total residual variance, to which $\sigma(\xi)$ is proportional, would be reduced, and secondly, these harmonics would no longer contribute at all to $\sigma(\xi)$. Thus it might be possible to restrict Stokes's integral to 10° or 15°, for although the uncertainty with a cap of radius 30° would not be altered very much, that with a cap of 10° or 15° would probably be reduced by about 0.5.

3.2. *The errors in the numerical evaluation of Stokes's integral*

The effects of uncertainties in the interpolation of the anomalies on the value of the integral will now be considered.

There is a difficulty in the use of the integral because it diverges at the origin, but this can be overcome by excluding the region within a small circle radius ψ about the origin from the integral, and treating it separately, for if $(\partial g/\partial x)_0$ is the mean gradient of gravity in the x -direction in this region, the deflexion in the x -direction may be shown to be

$$\frac{a\psi}{2g} \left(\frac{\partial g}{\partial x} \right)_0$$

(see, for example, Sollins 1947).

Suppose then that Stokes's integral extends from say 0.05 to ψ_0 , and divide the region by radii and concentric circle into compartments such that a mean anomaly of 1 mgal in each makes a constant contribution to the radial deflexion. Such a division is used in § 5, and in it an anomaly of 1 mgal in each compartment contributes 10^{-3} sec. to the radial deflexion. Since the radial divisions are at every 10°

the concentric circles are drawn with radii ψ_1 and ψ_2 such that $\int_{\psi_1}^{\psi_2} \frac{\partial f}{\partial \psi} \sin \psi d\psi$ is 0.171. If the mean anomaly in a compartment has a standard deviation σ mgal, the contribution to the standard deviation of the deflexion is

$$\frac{2\pi}{36} \frac{\sigma}{2\pi g} \left(\int_{\psi_1}^{\psi_2} \frac{\partial f}{\partial \psi} \sin \psi d\psi \right) \cos \bar{\alpha},$$

where $\bar{\alpha}$ is the angle between the direction in which the deflexion is computed and the direction of the centre of the compartment.*

If the means of the compartments have independent uncertainty (that is, systematic errors, such as those due to base station errors, are excluded), the variance of the deflexion is

$$\frac{1}{(36)^2} \left(\frac{\sigma}{g} \right)^2 (0.171)^2 \Sigma (\cos^2 \bar{\alpha}).$$

But $\Sigma \cos^2 \bar{\alpha}$ is nearly $\int_0^{2\pi} \cos^2 \alpha d\alpha$, that is, π . Hence the standard deviation of the calculated deflexion is $\frac{0.171}{36} \sqrt{\pi} \frac{\sigma}{g}$ for each concentric zone.

* If the anomalies were randomly distributed and the observations uniformly distributed, σ would be proportional to $1/\psi^{\frac{1}{2}}$; since neither of these conditions is fulfilled, σ has arbitrarily been taken to be constant.

Table 3 shows the number N of concentric circles for different outer limits of the integral when each compartment contributes 10^{-3} sec. per mgal, together with the corresponding standard deviation of the deflexion in meridian or prime vertical.

TABLE 3. ERRORS IN THE EVALUATION OF STOKES'S INTEGRAL

outer radius	number of circles	standard deviation of deflexion $\sigma \times$
5°	28	0°0094
10°	33	0°0101
15°	36	0°0106

The radii are at every 10°, σ , the standard deviation of the mean anomaly in a compartment, is in milligals.

Even if σ is as large as 10 mgals, and it is more likely to be about 5 mgals, the uncertainty of the deflexion is much less than that entailed in the estimation or neglect of the contributions of the rest of the world, and so much larger compartments could be used; for calculations of the difference between neighbouring stations, this size of compartment is found convenient (see § 5).

The contributions from the innermost region

This is the region in which the gradient of gravity is used to avoid the divergence of the integral.

If σ_g is the standard deviation of the gradient $\partial g/\partial x$ or $\partial g/\partial y$, the standard deviation of the deflexion is

$$\frac{a\psi}{2g} \sigma_g.$$

For representative values, take $a\psi = 5$ km., $g = 980$ cm./sec.², $\partial g/\partial x = 1$ mgal per km. Then the deflexion is approximately 0°5. There is very little idea of what σ_g may be; but the uncertainty in the deflexion is probably of the same order as that in the difference of two deflexions due to the neglect of the rest of the world outside a given small cap (§ 4.1).

3.3. Summary of errors of calculations of absolute deflexions

Let ψ_0 be the radius of the cap to which Stokes's integral is limited. The σ of § 3.2 is taken to be 10 mgals and σ_g to be 0.2 mgal.per km. Table 4 shows the estimates of the contributions to the standard deviations of the calculated deflexion.

TABLE 4

	ψ_0		
	10°	15°	20°
neglect of world outside ψ_0	1°6	1°1	0°7
evaluation of Stokes's integral from 5 km. to ψ_0	0°10	0°11	0°11
evaluation of Stokes's integral within 5 km.	0°1	0°1	0°1
total standard deviation	1°6	1°1	0°71

4. THE CALCULATION OF DIFFERENCES OF DEFLEXIONS AT NEIGHBOURING POINTS

It is very interesting to determine the errors likely in the difference of deflexions calculated at neighbouring points, for there are several applications for such differences and it may be possible to restrict Stokes's integral to very small caps about the points. But first, it is desirable to have some idea of the magnitude of the differences. The τ_1 and τ_2 variances are useless for this purpose, because if the points are only 1° or 2° apart, the deflexions due to harmonics of much higher order than the 18th will be nearly identical. It is the very local variations which cause the differences between points close together, and for them there are no data. Experience, however, indicates that the differences may often be two or three seconds and exceptionally are more.* This statement is based on the calculations made in § 6, on the deflexions found in the triangulation of Great Britain (Clarke 1858) and on some results in the United States (Rice 1949).

4.1. Error due to restriction of Stokes's integral to local caps

We now estimate the error in the calculated difference of deflexion between two adjacent points when the integration round each extends only to a distance ψ_0 .

Suppose the distance between the two points is β and the azimuth of the second from the first is α . Consider an anomaly Δg distance ψ , azimuth χ from the first point and $\psi + \delta\psi$, $\chi + \delta\chi$ from the second. If, $\beta \ll 1$,

$$\delta\psi \doteq \beta \cos(\alpha - \chi),$$

and

$$\delta\chi \doteq \frac{\beta \sin(\alpha - \chi)}{\sin \psi} \quad \text{provided } \beta \ll \psi.$$

Note that β/ψ may be quite large, for the expression given for $\delta\chi$ is correct if $\psi + \delta\psi$ equals ψ and the largest values of $\delta\chi$ occur when $\psi + \delta\psi$ is nearly equal to ψ ; when $\psi + \delta\psi$ is most different from ψ , i.e. when $(\alpha - \chi)$ is 0 or π , $\delta\chi$ is zero.

The contribution, I_1 , by anomalies between ψ_0 and π to the meridional deflexion at the first point is

$$\int_0^{2\pi} d\chi \int_{\psi_0}^{\pi} \Delta g \frac{\partial f}{\partial \psi} \sin \psi \cos \chi d\psi,$$

the factor $\frac{1}{2}\pi g$ being omitted.

To obtain I_2 , ψ must be varied to $(\psi + \delta\psi)$ and χ to $(\chi + \delta\chi)$, and the limits of integration must be altered so that $(\psi + \delta\psi)$ ranges from ψ_0 to π , and $(\chi + \delta\chi)$ from 0 to 2π .

Put

$$F = \Delta g \frac{\partial f}{\partial \psi} \sin \psi \cos \chi.$$

$$\begin{aligned} \text{Then } \delta I = I_2 - I_1 &= \int_0^{2\pi} (F(\psi_0) - F(\pi)) \delta\psi d\chi + \int_{\psi_0}^{\pi} (F(0) - F(2\pi)) \delta\chi d\psi \\ &\quad + \int_0^{2\pi} \int_{\psi_0}^{\pi} \left(\frac{\partial F}{\partial \psi} \delta\psi + \frac{\partial F}{\partial \chi} \delta\chi \right) d\psi d\chi. \end{aligned}$$

The second integral is zero since

$$F(\psi, 0) = F(\psi, 2\pi).$$

* In Cyprus, there are differences of nearly $1'$ for two stations about 1° apart.

Assume, as in § 3, that Δg is expanded about the first point in spherical harmonics, so that

$$F = \sum a_{nm} \cos m\chi P_n^m(\cos \psi) \frac{\partial f}{\partial \psi} \cos \chi \sin \psi$$

with a similar sum involving $b_{nm} \sin m\chi$.

The first integral then reduces to

$$\pi a_n \beta \cos \alpha \left[P_n \frac{\partial f}{\partial \psi} \sin \psi \right]_{\psi_0}^{\psi_1}.$$

In the last integral we have to consider $\partial F / \partial \psi$ and $\partial F / \partial \chi$ in which the differentiations refer to variations of the second point only and hence the Δg term is unaffected, i.e.

$$\frac{\partial F}{\partial \psi} = \sum a_{nm} \cos m\chi P_n^m \frac{\partial}{\partial \psi} \left(\frac{\partial f}{\partial \psi} \cos \chi \sin \psi \right) + \dots$$

The integral can be reduced, by integration by parts, to

$$\left[P_n \frac{\partial f}{\partial \psi} \sin \psi \right]_{\psi_0}^{\psi_1} + \int_{\psi_0}^{\psi_1} (P_n - P_n^1 \sin \psi) \frac{\partial f}{\partial \psi} d\psi.$$

Adding all the terms of δI , the contribution from the n th harmonic is found to be

$$\pi a_n \beta \cos \alpha \int_{\psi_0}^{\psi_1} (P_n - P_n^1 \sin \psi) \frac{\partial f}{\partial \psi} d\psi.$$

An approximate estimate of the integral can be obtained as follows:

For small ψ

$$P_n - P_n^1 \sin \psi \doteq 1.*$$

Further, $\partial f / \partial \psi$ decreases rapidly as ψ increases, so that much the largest contributions to the integral come from ranges in which ψ is close to ψ_0 . Accordingly, the integral may be replaced, approximately, by $\int_{\psi_0}^{\psi_1} \frac{\partial f}{\partial \psi} d\psi$, that is, $f(\pi) - f(\psi_0)$.

The uncertainty of the calculated difference of deflexion is accordingly

$$\frac{\pi \beta \cos \alpha (f(\pi) - f(\psi_0))}{2\pi g} (\sum a_n^2)^{\frac{1}{2}}.$$

On making the same assumptions about the distribution of the coefficients of the harmonics as in § 3, the variance due to the term $a_n P_n(\cos \theta)$ is found to be $a_n^2 / (2n + 1)$, and this must be equal to $38 / (2n + 1)$ (mgal)².

Therefore

$$a_n = 6.16 \text{ mgals.}$$

Since it is assumed that harmonics from $n = 4$ to $n = 18$ contribute to the error,

$$(\sum a_n^2)^{\frac{1}{2}} = 6.16 \sqrt{15} = 23.88 \text{ mgals.}$$

The standard deviation of the difference of the deflexions is therefore

$$\frac{23.88}{2g} \{f(\pi) - f(\psi_0)\} \beta \cos \alpha.$$

* For small ψ , $P_n \sim 1 - \frac{n\psi^2}{2}$, $P_n^1 \sim -n\psi$ and therefore $P_n - P_n^1 \sin \psi \sim 1 + \frac{n\psi^2}{2}$.

(For deflexions in the prime vertical, $\sin \alpha$ replaces $\cos \alpha$.) In seconds of arc, the standard deviation is $2.5 \{f(\pi) - f(\psi_0)\} \beta \cos \alpha$.

Table 5 gives values of the standard deviation for different values of β and ψ_0 , $\cos \alpha$ being 1.

TABLE 5. VALUES OF $2.5 \{f(\pi) - f(\psi_0)\} \beta$ IN SECONDS*

$\beta \backslash \psi_0$	0.5	1°	1.5	2°
5°	0.27	0.54	0.81	1.08
7.5	0.17	0.34	0.51	0.68
10°	0.12	0.24	0.36	0.48

* The values of f are taken from Lambert & Darling's tables (1936). For ψ_0 equal to 5°, it is desirable to include harmonics higher than the 18th, but no separate estimate of their variance exists.

Because of the approximations made in deducing the formula, the values for the larger ψ_0 and β must be rather rough, particularly when ψ_0 is not much greater than β .

These uncertainties will be almost unaffected by the determination of higher harmonics of gravity, except so far as the $(\tau_1 + \tau_2)$ variance is reduced, for all the harmonics contribute almost equally to the uncertainties.

4.2. *The standard deviation of the calculated difference of two deflexions*

The errors of the numerical calculations are the same as for an absolute deflexion. The contributions to the standard deviation are given in table 6.

TABLE 6. ERRORS OF A DIFFERENCE OF DEFLEXIONS AT STATIONS 1° APART

	ψ_0	
	5°	10°
neglect of world outside ψ_0	0.54	0.24
evaluation of Stokes's integral from 5 km. to ψ_0	0.09	0.10
evaluation of Stokes's integral within 5 km.	0.1	0.1
total standard deviation	0.56	0.28

5. SOME CALCULATIONS OF DEFLEXIONS

The calculations which have been made use the observations of gravity in the British Isles, the surrounding waters, and western Europe, and were undertaken with the objects of comparing the errors obtained with the uncertainties predicted in the preceding sections, and of discovering how suitable the present data are for these calculations. Thus, differences of deflexions at neighbouring places were considered, since these are equal to the differences of the astronomical and geodetic amplitudes between the places, and are therefore known or can be found fairly readily. The places chosen in this country are Greenwich, Herstmonceux and Southampton; the differences between Greenwich and Southampton are already known (Clarke 1858), and those between Herstmonceux and these places will be known when the new Royal Observatory is properly established there. Other places

at which there are astronomical observations of course exist, such as Cambridge, Arbury Hill, Liverpool, or Clifton (Yorks), but the gravity data are for one reason or another not yet suitable for calculations at these places. The calculations in this country were made to compare the actual errors with the predicted uncertainty; in addition, calculations were made at Bayeux, since, if the differences of the deflexions here and at Southampton are known, astronomical observations at the two places provide a connexion between the French and English geodetic triangulations additional to that across the Straits of Dover. This may help to check the relative azimuths of the two portions of the Anglo-Gallic arc of meridian, an error in which could affect the contribution of this arc in the determination of the earth's figure.

5.1. *The gravity data*

These are the free-air anomalies in western Europe and the British Isles, of which a map is shown in figure 1. The individual stations cannot be shown except where they are very sparse; elsewhere lines of equal anomaly are drawn at intervals of 10 mgals. In areas where there are no observations, hypothetical values, indicated by broken figures, have had to be used; they are based on surrounding observations where possible. The largest blank area is the North Sea, in which the anomaly has been assumed to be zero. Most of the surveys in and around the British Isles have been mentioned in the introduction (White 1949; Cook & Thirlaway 1950; Cook, Murphy & Thirlaway 1950; Bullard & Jolly 1936; Browne & Cooper 1950; Lejay 1947). Papers by Ackerl (1932) and Meinesz (1923) have also been consulted, and the Directors of the Anglo-Iranian Oil Company and M. Lejay have most generously allowed me to see unpublished material.

All values of gravity are on the Potsdam standard, and the anomalies are calculated with the International formula (as tabulated by the U.S. Coast and Geodetic Survey).

Base station values

There is some uncertainty about the relative values of anomalies in the different countries because of the uncertainties of the comparison of the national base stations to which they are referred. To estimate the magnitude of the error, suppose that all the values of gravity to the north of the station are 1 mgal high compared to those south of the station. Then if the limits of Stokes's integral are $0^{\circ}05$ and 5° , which are the limits used below, the deflexion will be in error by about $0''.3$.

The base station values actually used are:

Cambridge	981.265 cm./sec. ²
Paris	980.944
de Bilt	981.268

On comparing these with the best values at present obtainable* the difference between Cambridge and Paris is found to be 2.9 mgal too small, and the difference between Cambridge and de Bilt to be 2.4 mgal too great. Corrections will be applied for these errors in the final stages of the calculations.

* Calculated by the author in a thesis for the Ph.D. degree submitted to the University of Cambridge, 1949.

5.2. Arrangement of the calculations

5.2.1. Stokes's integral

Stokes's integral is used for the region from $0^{\circ}05'$ to 5° . The region exterior to this is neglected, and the interior region, where Stokes's integral diverges, is treated specially.

The free-air anomalies were plotted on a one in one million map of the whole area shown in figure 1 and on $\frac{1}{4}$ in. to the mile maps of the immediate neighbourhood of each station. Templates were drawn for these two scales which divided the whole



FIGURE 1. Free-air anomalies in north-west Europe. ● submarine pendulum stations (free-air anomalies in mgals); ○ places at which deflexions are computed; —, lines of equal free-air anomalies (mgals); the large figures in broken line indicate the anomalies assumed in regions without observations.

TABLE 7. SOUTHAMPTON: MEAN FREE-AIR ANOMALIES (MGALS) IN COMPARTMENTS ON $\frac{1}{4}$ IN. MAP

radius (km.)	E					W				
angle	0°	30°	60°	90°	120°	150°	180°	210°	240°	270°
5.56	-25	-25	-25	-25	-26	-27	-27	-27	-27	-27
11.00	-23	-23	-23	-23	-26	-26	-26	-26	-26	-26
21.78	-15	-15	-17	-17	-20	-10	-10	-10	-10	-10
43.00	-7	-10	-10	-15	-20	-15	-15	-15	-15	-15
84.45	-70	-73	-75	-80	-91	-77	-77	-77	-77	-77
total										
radius (km.)	S					N				
angle	180°	210°	240°	270°	300°	330°	360°	330°	300°	270°
5.56	-27	-27	-27	-28	-28	-27	-27	-27	-27	-27
11.00	-26	-26	-25	-20	-25	-25	-25	-25	-25	-25
21.78	-10	-15	-17	-20	-20	-17	-15	-15	-15	-15
43.00	-17	-10	-5	-5	-2	-10	-5	-5	-5	-5
84.45	-80	-78	-74	-78	-75	-77	-71	-71	-71	-71
total										

Each compartment contributes 4×10^{-3} sec. per mgal radially.

area, by radii at each 10° and by concentric circles, into compartments each of which contributed to the deflexion 0.001 radially on the 1 m. scale and 0.004 on the ¼ in. scale for a mean anomaly of 1 mgal. The radii of the circles are the following (in kilometres):

on the ¼ in. scale:

5.56 (0°05), 11.00, 21.78, 43.00, 84.45;

on the 1 m. scale:

84.45, 99.85, 117.98, 139.36, 164.31, 193.60, 227.84, 267.73, 314.07, 367.70, 429.52, 500.44, 581.38 (5°23).

These compartments are convenient for the estimation of mean values on the particular scales.

The contributions for all compartments between a pair of radii are added together, and the sum, multiplied by the sine or cosine of the angle between the mean direction of the sector and the meridian, gives the east and north components of the deflexion due to this sector. As an example, the details for the ¼ in. map for Southampton are given in tables 7 and 8. The first shows the mean anomalies in each compartment and the sector sums and the second the calculation of the components of the deflexion.

TABLE 8. SOUTHAMPTON: CONTRIBUTIONS FROM ¼ IN. MAP-COMPONENTS OF DEFLEXIONS

sector	contribution (east positive)	cos	north component	sin	east component
0°					
10°	+ 10	0.996	9.96	0.087	0.87
20°	+ 5	0.966	4.83	0.259	1.30
30°	+ 13	0.906	11.78	0.423	5.50
40°	+ 7	0.819	5.73	0.574	4.02
50°	- 1	0.707	- 0.71	0.707	- 0.71
60°	- 1	0.574	- 0.57	0.819	- 0.82
70°	+ 3	0.423	1.27	0.906	2.72
80°	- 7	0.259	- 1.81	0.966	- 6.76
90°	- 7	0.087	- 0.61	0.996	- 6.97
100°	- 15	- 0.087	1.31	0.996	- 14.94
110°	- 16	- 0.259	4.14	0.966	- 15.46
120°	- 14	- 0.423	5.92	0.906	- 12.68
130°	- 4	- 0.574	2.30	0.819	- 3.28
140°	- 1	- 0.707	0.71	0.707	- 0.71
150°	- 3	- 0.819	2.46	0.574	- 1.72
160°	- 6	- 0.906	5.44	0.423	- 2.54
170°	- 14	- 0.966	13.52	0.259	- 3.63
180°	- 11	- 0.996	10.96	0.087	- 0.96

1 mgal contributes 0.004 per compartment.
Hence: total north component = 0.306.
total east component = - 0.226.

5.2.2. The local contributions

For none of the four places are there sufficient observations in the region out to 5.56 km. to enable the local deflexion to be found from the free-air anomalies by estimation of the gradient at the point of computation.

The attraction of the topography has therefore been calculated separately by Clarke's method (Clarke 1858), and to this has been added the effect of the gradient of the Bouguer anomaly. The attraction of the topography is proportional to the density of the rock; this has been taken to be 2.0 g./cm.³.

The maps used in calculating the topographic effect were, in England, the 6 in. to 1 mile map out to 5000 ft. and the 1 in. to 1 mile map for the remainder.

5.3. Results

The results are given in table 9.

The co-ordinates refer to the meridian circle at Greenwich and the Ordnance Survey Office at Southampton. The site of the zenith telescope at Herstmonceux was unknown when the calculations were made, and the exact site of the trigonometrical station at Bayeux was unknown and the calculations for the region out to 5000 ft. were therefore not made for these places.

TABLE 9. STATIONS AND COMPONENTS OF DEFLEXIONS

station and co-ordinates	region	data	north deflexion	east deflexion	resultant	
					magni- tude	direction from E by N
Greenwich lat. 51° 28' 38" N long. 0° 00' 00" E	581.38 to 84.45 km.	free-air	1"31	-0"98	1"64	323°4
		anomalies				
	84.45 to 5.56 km.	free-air	1"32	-0"59	1"45	336°0
		anomalies				
	within 5.56 km.	topography	-0"49	0"35	0"60	35°5
		Bouguer	0"17	-0"12	0"21	324°8
		anomalies				
	total		2"32	-1"33	2"67	330°2
Herstmonceux lat. 50° 52' 15" N long. 0° 20' 5 E	581.38 to 84.45 km.	free-air	0"75	-0"48	0"89	327°7
		anomalies				
	84.45 to 5.56 km.	free-air	1"69	1"21	2"07	35°6
		anomalies				
	within 5.56 km.	topography	0"18	0"03	0"18	9°5
		Bouguer	0"17	0"12	0"21	35°2
		anomalies				
	total		2"79	0"88	2"93	17°5
Southampton lat. 50° 54' 49" N long. 01° 24' 06" W	581.38 to 84.45 km.	free-air	1"00	-2"53	2"72	338°4
		anomalies				
	84.45 to 5.56 km.	free-air	0"31	-0"23	0"38	323°6
		anomalies				
	within 5.56 km.	topography	0"19	0"06	0"20	17°6
		Bouguer	0"00	0"14	0"14	90°0
		anomalies				
	total		1"50	2"56	2"97	329°6
Bayeux lat. 49° 17' N long. 00° 43' W	581.38 to 84.45 km.	free-air	-0"92	-1"78	2"01	207°4
		anomalies				
	84.45 to 5.56 km.	free-air	-3"86	0"12	3"86	107°8
		anomalies				
	within 5.56 km.		estimated zero on insufficient data			
	total		-4"78	-1"66	5"06	199°1

Effect of base station errors

Bayeux is the only station for which the errors in the base station values have a significant effect and they have a maximum effect there because all stations to the

south are based on Paris and all to the north, including the English Channel, are based on Cambridge. According to the results given above (§ 5.1) the northward deflexion at Bayeux should be increased by 0.3×2.9 to -3.91 .

Regions with few observations

These are the North Sea, affecting especially Greenwich and Herstmonceux, and the Channel Islands, the west coast of France and the Bay of Biscay, affecting Bayeux. In the Weald of Kent a free-air anomaly map was constructed by adding the attraction of the topography to Bouguer anomalies based on scattered stations.

The regions with few observations are rather far from Southampton and the calculations here are the most reliable of the four.

5.4. *Comparison of observed and calculated differences of deflexion*

The only comparison which can be made is for the meridional deflexions at Greenwich and Southampton.

The astronomical latitudes are:

Greenwich: $51^{\circ} 28' 38''.2$ (*Nautical Almanack* 1938)

Southampton: $50^{\circ} 54' 46''.97$ (Clarke 1858)

and the amplitude is therefore

$$33' 51''.2.$$

The geodetic amplitude between the parallels of Greenwich and Southampton is $33' 50''.7$ using the elements of the International figure.

The difference between the astronomical and geodetic amplitudes is thus $0''.5$.

From the deflexions given in table 9 the difference should be $-0''.8$.

The difference of the observed and calculated deflexions is therefore $1''.3$.

The actual distance of Southampton from Greenwich is $1^{\circ} 2'$ and the azimuth of Greenwich from Southampton is $56^{\circ} 49'$; the standard deviation of the difference of the deflexion in the meridian is therefore, according to table 5, $0''.54 \cos 56^{\circ} 49'$, that is, $0''.3$. The error of the calculated difference of deflexions is thus four times the standard deviation.

Errors in the geodetic triangulation can be excluded completely. The other sources of error are then:

- the astronomical latitude at Southampton;
- the radius of curvature of the earth;
- the calculated deflexions.

The astronomical observations at Southampton extended over about 10 months, and were made with Airy's zenith sector. The 14-monthly component of the variation of latitude (total amplitude about $0''.2$) will thus not be completely eliminated, and further, the corrections to be applied to work with Airy's sector have given rise to some discussion (Clarke 1858). The result for Southampton may therefore be in error, but almost certainly by less than $0''.5$.

Since the geodetic amplitude between the stations is obtained by triangulation it is unaffected to first-order by deformations of the geoid at intermediate points and depends therefore on the base measurements, the triangulation and the mean radius of curvature at this latitude, and the error should be less than $0''.1$.

It seems unlikely that these errors could contribute more than 0.5 to the discrepancy of 1.3, and the influence of the anomalies in the unsurveyed regions, particularly the North Sea, must be considered. An anomaly of -10 mgal in this area would reduce the discrepancy of the deflexions by 0.5, and the remaining 0.8 would be then about twice the standard deviation of the difference of deflexions.

Some further indication of the effect of the North Sea may be obtained by comparing the deflexions found above with those determined by Clarke (1858) from the least squares adjustment of all the geodetic and astronomical observations in the British Isles:

	from gravity	Clarke
Greenwich: meridian	2.32	1.48
prime vertical	1.33	-0.40
Southampton: meridian	1.50	1.90

The agreement at Southampton is satisfactory; at Greenwich the meridian difference indicates a large area of negative free-air anomalies to the north. These results therefore make it very probable that the free-air anomaly over much of the North Sea is less than -10 mgals. The Bouguer anomalies of about -10 mgals on the east coast of Yorkshire (White 1949) are consistent with such anomalies over the North Sea. As more gravity measurements are made on land it should be possible to make calculations for other astronomical stations and to infer the anomalies in the North Sea in more detail.

These results suggest that it may be possible to infer the distribution of gravity anomalies in areas inaccessible to direct observation by measuring deflexions of the vertical round the margins of the area and comparing them with those calculated from gravity anomalies outside the area.

6. CONCLUSIONS

The studies of the errors of the calculations of deflexions of the vertical show that quite useful calculations may be made by using Jeffreys's determination of the lower harmonics of the free-air anomalies to give the contribution to the deflexion from most of the world, the local part being found by using Stokes's integral over a cap of 10° to 15° radius. Although at present the uncertainties of the deflexions so calculated are rather large, they would be considerably reduced if gravity measurements were made in the South Pacific. Any other procedure would give larger uncertainties and would not be so convenient for computation, since the one discussed here reduces the numerical work for each station to a minimum. If only differences between neighbouring points are required, the effects of the low harmonics of gravity can be neglected and Stokes's integral can be restricted to between 5° and 10° from the station.

Until the results of the gravity surveys at present in progress in England are available, it would not be worth while to make more calculations of deflexions in this country. When the anomalies are better known in the Midlands, a computation for Oxford Observatory would be useful, for this is so situated that a comparison with Southampton would not be too much affected by the North Sea. With more

calculations and comparisons it should be possible to estimate the anomalies in the North Sea even if submarine gravity observations there remain, as at present, impractical.

The calculation of deflexions of the vertical was suggested to me by Mr B. C. Browne, and I am indebted to him and to Professor Jeffreys and Dr de Graaff Hunter for discussions on theoretical points. I am grateful to Miss C. Munford for the computation of table 1.

I wish to thank the Department of Scientific and Industrial Research for a maintenance grant.

APPENDIX 1

The following calculations show what errors may occur in hilly regions because the tangent plane to the equipotential surface at an external point is not parallel to the tangent plane to the geoid.

The change of deflexion with height is calculated for a set of parallel sinusoidal ridges which might represent country such as Salisbury Plain or the Lake District.

The sinusoidal topography

It is assumed that the region may be represented as a set of infinite ridges and valleys of mean height a whose gravity field is equivalent to that of a surface layer of density $A \cos px$ at height a . The problem is treated in two dimensions—height h above sea-level and horizontal distance x perpendicular to the axis of the ridges and valleys.

The horizontal component of gravity is

$$2\pi f A e^{-p(h-a)} \sin px,$$

and if g is the total attraction of gravity due to the whole earth, the deflexion of the vertical is

$$\frac{2\pi f}{g} A e^{-p(h-a)} \sin px.$$

Let $2b$ be the height of the crests above the valleys and ρ the density of the rocks. Then the deflexion is

$$\frac{2\pi f \rho b}{g} e^{-p(h-a)} \sin px.$$

(The deflexion vanishes on the crests and in the troughs.)

In the following numerical examples g is taken as 10^6 mgals and ρ as 2.5 g./cm.³.

Example 1

Undulating plateau, mean height $a = 500$ ft. amplitude, $b = 200$ ft.

Wave-length $= 2\pi/\rho = 5000$ ft.

(This is somewhat like Salisbury Plain.)

The deflexion at the mean height is about $0''.5$ and at sea-level, $0''.25$.

Both the deflexion and the error are thus quite small. If the wave-length were 10,000 ft., the deflexion at sea level would be about $0''.36$.

Example 2

Mountainous country: $a = 1500$ ft.

$b = 1500$ ft.

Wave-length, 10,000 ft.

(Somewhat like the Lake District.)

Deflexion at the mean height is about $9''.6$.

Deflexion at sea-level is about $3''.5$.

These two examples will indicate the magnitude of the error.

Suppose that the angle between the normals to the external equipotential surface and the geoid is to be less than $0''.25$, a satisfactory limit in relation to the other errors involved in the calculation of deflexions.

Then for this 'sinusoidal' topography of density 2.5 g./cm.³, the condition is

$$0.64b\{1 - \exp(-2\pi a/\lambda)\} < 0.25,$$

where b is in hundreds of feet.

$2\pi a/\lambda$ must in general be small, so that, roughly,

$$\lambda/a > 5.12\pi b.$$

REFERENCES

- Ackerl, F. 1932 *S.B. Akad. Wiss. Wien (Math.-Phys.)*, **141**, 303-439.
- Browne, B. C. & Cooper, R. I. B. 1950 *Phil. Trans. A*, **242**, 243-310.
- Bullard, E. C. & Jolly, H. L. P. 1936 *Mon. Not. R. Astr. Soc., Geophys. Suppl.*, **3**, 443-477.
- Clarke, A. R. 1858 *Account of the principal triangulation*. London: Ordnance Survey.
- Cook, A. H. & Thirlaway, H. I. S. 1950 *Rep. XVIII Int. Geol. Congr.* London, part v.
- Cook, A. H., Murphy, T. & Thirlaway, H. I. S. 1950 *Geophys. Mem. Dublin Inst. Adv. Stud.* no. 2 (3 parts).
- Darling, F. W. 1948 *Bull. Geod.* no. 9, pp. 199-240.
- Hunter, J. de Graaff 1935 *Phil. Trans. A*, **234**, 377-431.
- Jeffreys, H. 1931 *Beitr. Geophys.* **31**, 378-386.
- Jeffreys, H. 1932 *Beitr. Geophys.* **36**, 206-211.
- Jeffreys, H. 1933 *Beitr. Geophys.* **39**, 374-377.
- Jeffreys, H. 1941 *Mon. Not. R. Astr. Soc., Geophys. Suppl.*, **5**, 1-22.
- Jeffreys, H. 1943 *Mon. Not. R. Astr. Soc., Geophys. Suppl.*, **5**, 55-66.
- Jeffreys, H. 1948 *Mon. Not. R. Astr. Soc., Geophys. Suppl.*, **5**, 219-247.
- Lambert, W. D. & Darling, F. W. 1936 *Spec. Publ. U.S. Cst Geod. Survey*, no. 199.
- Lejay, P. 1947 *C.R. Acad. Sci., Paris*, **224**, 1086 and **225**, 10.
- Meinesz, F. A. Vening 1923 *Observations de pendule dans les Pays-Bas, 1913-1921. Publ. Comm. géod. néerland.* Delft: J. Waltman.
- Meinesz, F. A. Vening 1928 *Proc. K. Akad. Wet. Amst.* **31**, 315.
- Meinesz, F. A. Vening 1941 *Gravity expeditions at sea, 1934-1939*, **3**. Delft: J. Waltman.
- Rice, D. A. 1949 *Trans. Amer. Geophys. Un.* **30**, 323-327.
- Sollins, A. D. 1947 *Bull. Geod.* no. 6, pp. 279-300.
- Steenwijk, Baron J. E. De Vos von 1948 *Plumb-line deflections and geoid in Eastern Indonesia. Publ. Comm. géod. néerland.* Delft: J. Waltman.
- Stokes, G. G. 1849 *Trans. Camb. Phil. Soc.* **8**, 672.
- Tanni, L. 1948 *Ann. Acad. Sci. Fenn. (ser. v)*, **3**. Geol.-Geogr., p. 16.
- White, P. H. N. 1949 *Quart. J. Geol. Soc. Lond.* **104**, 339-364.

Fermi-Dirac functions of integral order

BY P. RHODES, PH.D., *Physics Department, University of Leeds*

(Communicated by E. C. Stoner, F.R.S.—Received 22 July 1950)

After a brief indication of the types of physical problem in which they arise, an account is given of methods of evaluation of the Fermi-Dirac functions, $F_n(\eta) = \int_0^\infty \{x^n / (e^{x-\eta} + 1)\} dx$, for positive integral values of n . The following relationship is derived:

$$F_n(+|\eta|) + (-)^{n+1} F_n(-|\eta|) = S_n(+|\eta|),$$

where
$$S_n(\eta) = \frac{\eta^{n+1}}{n+1} \left\{ 1 + \sum_{r=1}^{\infty} 2(n+1) n \dots (n-2r+2) (1-2^{1-2r}) \zeta(2r) \eta^{-2r} \right\};$$

and the expressions for $S_n(\eta)$ are tabulated for $n = 1, 2, 3, 4$. A series suitable for the evaluation of $F_n(-|\eta|)$ to any required accuracy is indicated; together with the derived relationship this provides a means by which $F_n(+|\eta|)$ may be computed to any required accuracy. To facilitate the use of the functions the values of $(1/n!) F_n(-|\eta|)$ for $n = 1, 2, 3, 4$ have been calculated and are tabulated to seven decimal places for $\eta = 0.0(0.1) 4.0$.

1. INTRODUCTION

The Fermi-Dirac functions, which arise in the theoretical treatment of assemblies of particles subject to Fermi-Dirac statistics, e.g. electrons, are defined by

$$F_q(\eta) = \int_0^\infty \frac{x^q dx}{e^{x-\eta} + 1}. \quad (1.1)$$

The physical significance of the parameters involved is indicated by the relations

$$x = \epsilon/kT, \quad \eta = \zeta/kT, \quad (1.2)$$

where ϵ and ζ are respectively the energy and chemical potential per particle. These functions cannot in general be expressed in terms of elementary functions. The evaluation of the functions corresponding to $q = \frac{1}{2}, \frac{3}{2}$ has been dealt with by McDougall & Stoner (1938, referred to as M.S.), who give tables and series covering effectively the whole range of possible values of η . These functions are required in the treatment of the equilibrium magnetic and thermal properties of assemblies of electrons having a standard energy distribution of states, i.e. for which the density of states, $dv/d\epsilon$, is of the form $dv/d\epsilon = C\epsilon^{\frac{1}{2}}$, where C is a constant. For a general discussion of the properties of the functions and of their applications reference may be made to M.S. and to Stoner (1939a).

In the treatment of transport effects, e.g. electrical and thermal conductivity, in assemblies of electrons distributed in a band of standard form, and in the treatment of the equilibrium properties of electrons distributed in a band of the form $dv/d\epsilon = C\epsilon^n$, where n is a positive integer (including zero), Fermi-Dirac functions in which q is a positive integer are required (cf. §2). Although the evaluation of these functions is much simpler than that of those treated in M.S. it does not appear to have been discussed in detail previously. In the present paper the methods of evaluation are

considered. A relation is derived between values of the functions with positive and negative arguments, and a convergent series is given for the evaluation of those with negative arguments (§3). For $q = 1, 2, 3, 4$ numerical tables are given for $F_q(-|\eta|)$, which together with the series expressions and derived relation cover effectively the whole range of possible values of η (§4).

2. PHYSICAL APPLICATIONS OF THE FUNCTIONS

Before considering the evaluation of the functions a brief indication will be given of some of their fields of application.

Transport effects

A lengthy discussion of this topic would be inappropriate here, and only the relevant results will be quoted. For further details reference must be made to some account of the Bloch treatment of transport effects in metals, e.g. Sommerfeld & Bethe (1933), Wilson (1936) (cf. also Rhodes 1950).

In the presence of an electric field, F , parallel to the z -axis, and a temperature gradient dT/dz , the steady-state electron momentum distribution function in a metal or semi-conductor may be written as

$$f(\mathbf{k}) = f_0(\epsilon) - k_z \mu(\epsilon) \frac{df_0}{d\epsilon}, \quad (2.1)$$

where k_z is the z -component of the electron wave-vector $\mathbf{k}$, and f_0 is the equilibrium, Fermi-Dirac, distribution function,

$$f_0(\epsilon) = \frac{1}{e^{x-\eta} + 1}, \quad (2.2)$$

in which x and η are defined by (1.2). The theoretical treatment shows that the function $\mu(\epsilon)$ may be expressed in the form (cf. Wilson 1936, pp. 215 et seq.)

$$\mu(\epsilon) = \alpha \left[\left\{ eF + T \frac{d}{dz} \left(\frac{\zeta}{T} \right) \right\} \mu_1(\epsilon) + \left(\frac{1}{T} \frac{dT}{dz} \right) \mu_2(\epsilon) \right], \quad (2.3)$$

where $-e$ is the electron charge, α is a function of the temperature and of the characteristics of the metal, but is independent of ϵ , and $\mu_1(\epsilon)$, $\mu_2(\epsilon)$ are determined by the solutions of two integral equations. In terms of these functions, the electrical and thermal current densities, J and w respectively, may be written (cf. Wilson 1937, p. 373)

$$J = M_1 e \left\{ eF + T \frac{d}{dz} \left(\frac{\zeta}{T} \right) \right\} + M_2 \frac{e}{T} \frac{dT}{dz}, \quad (2.4)$$

$$w = M_2 \left\{ -eF - T \frac{d}{dz} \left(\frac{\zeta}{T} \right) \right\} - M_3 \frac{1}{T} \frac{dT}{dz}, \quad (2.5)$$

where

$$\left. \begin{aligned} M_1 &= -\beta \int_0^\infty \epsilon^\dagger \mu_1(\epsilon) f'_0 d\epsilon, \\ M_2 &= -\beta \int_0^\infty \epsilon^\dagger \mu_2(\epsilon) f'_0 d\epsilon, \\ M_3 &= -\beta \int_0^\infty \epsilon^\dagger \mu_2(\epsilon) f'_0 d\epsilon, \\ f'_0 &= df_0/d\epsilon, \end{aligned} \right\} \quad (2.6)$$

and β is a function of the temperature and of the characteristics of the metal. It is readily shown from (2.4), (2.5) that the electrical conductivity, σ , thermal conductivity, κ , and thermoelectric power, ϕ , are given by

$$\left. \begin{aligned} \sigma &= e^2 M_1, \quad \kappa = \frac{M_1 M_3 - M_2^2}{M_1 T}, \\ \phi &= -\frac{1}{e} \left(\frac{M_2}{M_1 T} - \frac{\zeta}{T} \right). \end{aligned} \right\} \tag{2.7}$$

The applications of the Fermi-Dirac functions may be illustrated by considering cases in which the functions $\mu_1(\epsilon)$, $\mu_2(\epsilon)$ take particularly simple forms. For metals at high temperatures (corresponding to $(T/\theta_D) \gg 1$, where θ_D is the Debye temperature)

$$\mu_n(\epsilon) = a\epsilon^{n+1/2}, \tag{2.8}$$

where a is independent of ϵ , and n is integral (in particular $n = 1, 2$). The corresponding expressions for the M_n from (2.6) become

$$M_n = -a\beta \int_0^\infty \epsilon^{n+2} f'_0 d\epsilon. \tag{2.9}$$

For semi-conductors (cf. Wilson 1936, p. 211)

$$\mu_n(\epsilon) = b\epsilon^{n-1/2}, \tag{2.10}$$

where b is independent of ϵ ; and from (2.6)

$$M_n = -b\beta \int_0^\infty \epsilon^n f'_0 d\epsilon. \tag{2.11}$$

Similarly, Mestel (1950) in a discussion of the thermal conductivity associated with electrons in dense stars obtains an expression for the steady-state momentum distribution function which is essentially equivalent to that given by (2.1), (2.3), (2.8), so the thermal conductivity could be expressed by the relations (2.7), (2.9).

The integrals in (2.9), (2.11) are members of the sequence

$$K_n(\eta) = - \int_0^\infty \epsilon^n f'_0 d\epsilon, \tag{2.12}$$

and these may be expressed in terms of the basic functions (1.1). Changing the variable in (2.12) to x and integrating by parts,

$$\begin{aligned} K_n(\eta) &= - \int_0^\infty (kT)^n x^n \frac{d}{dx} \left(\frac{1}{e^{x-\eta} + 1} \right) dx \\ &= -(kT)^n \left\{ \left[\frac{x^n}{e^{x-\eta} + 1} \right]_0^\infty - n \int_0^\infty \frac{x^{n-1} dx}{e^{x-\eta} + 1} \right\} \\ &= n(kT)^n F_{n-1}(\eta). \end{aligned} \tag{2.13}$$

It can be seen from (2.7), (2.9), (2.11) that the values of K_n of particular interest are those for which $n = 1(1) 5$, corresponding to values of F_n with $n = 0(1) 4$. In illustration, the expression for the thermal conductivity in a metal may be written from (2.7), (2.9), (2.13) as

$$\kappa = a\beta k^5 T^4 \frac{(15F_2 F_4 - 16F_3^2)}{3F_2}, \tag{2.14}$$

and similar expressions may be written down for the other transport effects.

In applications of (2.7), (2.9) to metals the integrals are usually evaluated by the 'low-temperature approximation', which is valid only for $\eta \gg 1$ (cf. §3). For semiconductors the required integrals are usually evaluated by the 'high-temperature approximation', corresponding to $-\eta \gg 1$ (cf. §3). However, these approximations are not always adequate. For example, Putley (1949) in a discussion of some experimental results for germanium found that the high-temperature approximation was not sufficiently accurate. Putley considered only the electrical conductivity, and in this case the relevant function, F_0 , could be expressed in terms of elementary functions, but the extension of the treatment to cover thermal conductivity or the thermo-electric effects would require further members of the sequence (1.1) (cf. (2.7), (2.11)). Similarly, in the consideration of the transport effects in stars values of the basic functions are required for values of η outside the ranges covered by the high- and low-temperature approximations (cf. Mestel 1950).

Rectangular and related band forms

Fermi-Dirac functions of integral order are also involved in the treatment of the equilibrium magnetic and thermal properties of particles (electrons or 'holes') distributed in an energy band of the form $d\nu/d\epsilon = C\epsilon^n$, where C is a constant and n is integral (including zero). The treatment in such a case is formally similar to that for particles in a band of standard form (see Stoner 1938, 1939*b*), but the functions $F_{\frac{1}{2}}, F_{\frac{3}{2}}$ are replaced by F_n, F_{n+1} . Up to the present the only such case which has been considered in detail is that of the magnetic properties of particles distributed in a rectangular band, corresponding to $n = 0$ (cf. Néel 1938). For this the relevant function, F_0 , may be evaluated directly as $\ln(1 + e^\eta)$. However, for the thermal properties in this case, and for both magnetic and thermal properties for band forms corresponding to larger values of n , further members of the sequence F_n are required. By means of the functions considered here (cf. §4), and those already dealt with in M.S., the properties of a wide variety of band forms could be treated with numerical precision.

3. EVALUATION OF $F_n(\eta)$ AND RELATION BETWEEN $F_n(\eta)$ AND $F_n(-\eta)$

In order to avoid excessive repetition it is convenient to write the functions to be considered in the form

$$F_n(\eta) = F_n = \int_0^\infty \frac{x^n dx}{e^{x-\eta} + 1}, \quad (3.1)$$

where n is a positive integer or zero, and $\eta \geq 0$. This convention is adopted throughout the remainder of the present paper unless otherwise stated.

The first member of the sequence (3.1) may be evaluated directly as

$$F_0(\eta) = \ln(1 + e^\eta). \quad (3.2)$$

Analytical expressions for the following members of the sequence cannot be obtained, but it is readily shown that the various members are related by

$$F'_n(\eta) = \frac{d}{d\eta} F_n(\eta) = nF_{n-1}(\eta), \quad (3.3)$$

or, in terms of the inverse relation,

$$F_n(\eta) = F_n(0) + n \int_0^\eta F_{n-1}(\eta) d\eta. \quad (3.4)$$

In principle, any member of the sequence may be derived from F_0 by successive applications of the relation (3.4), since, as indicated below, the $F_n(0)$ are directly expressible in terms of Riemann zeta functions. In practice this would require a large number of numerical integrations to determine the values of F_n corresponding to the larger values of n , and an alternative method, based on the relation which may be shown to exist between $F_n(\eta)$ and $F_n(-\eta)$, has been developed. Before considering this relation the method of evaluation of $F_n(-\eta)$ by means of a convergent infinite series must be indicated.

Series for $F_n(-\eta)$

For zero and negative values of the argument of F_n it may be shown (cf. M.S. p. 71) that

$$F_n(-\eta) = n! \sum_{r=1}^{\infty} (-)^{r+1} \frac{e^{-r\eta}}{r^{n+1}}. \quad (3.5)$$

$$\text{In particular, for } \eta = 0, \quad F_n(0) = n!(1 - 2^{-n}) \zeta(n+1), \quad (3.6)$$

where $\zeta(n+1)$ is a Riemann function. By means of the series (3.5) $F_n(-\eta)$ may be computed to any desired degree of accuracy, although the number of terms required may be very large for small values of η (cf. §4). It is the first term of this series which is usually referred to as the 'high-temperature approximation' to the function.

Polynomial approximation and relation between $F_n(\eta)$ and $F_n(-\eta)$

For $\eta \gg 1$ an 'asymptotic series' may be developed for $F_n(\eta)$ (cf. M.S. p. 81; Nordheim 1934). Application of the general formula for the series to integral values of n shows that in this case the series terminates after a finite number of terms, so that it is not in fact an asymptotic series but a polynomial in η . In examining the possibility of representing $F_0(\eta)$ by a convergent series, so as to avoid the necessity of repeated numerical integration in building up the following members of the sequence, a relation between $F_n(\eta)$, $F_n(-\eta)$ and the polynomial approximation became apparent. This relation, which allows $F_n(\eta)$ to be computed directly from $F_n(-\eta)$, or vice versa, may be derived by an extension of the analysis involved in the derivation of the polynomial approximation, as follows.

Consider the functions

$$I_n(\eta) = - \int_0^\infty x^n \frac{df_0}{dx} dx = \int_0^\infty \frac{x^n dx}{(e^{x-\eta} + 1)(e^{-x+\eta} + 1)}. \quad (3.7)$$

It is readily shown by an analysis similar to that used in deriving (2.13) that these functions are related to the Fermi-Dirac functions by

$$I_n(\eta) = nF_{n-1}(\eta). \quad (3.8)$$

Substituting $p = x - \eta$ in (3.7),

$$\begin{aligned} I_n(\eta) &= \int_{-\eta}^\infty \frac{(\eta + p)^n dp}{(e^p + 1)(e^{-p} + 1)} \\ &= \int_{-\infty}^\infty \frac{(\eta + p)^n dp}{(e^p + 1)(e^{-p} + 1)} - \int_{-\infty}^{-\eta} \frac{(\eta + p)^n dp}{(e^p + 1)(e^{-p} + 1)}. \end{aligned} \quad (3.9)$$

The first of these two integrals gives the polynomial approximation for I_n , and hence for nF_{n-1} , and it can be seen that the error introduced by neglect of the second term is of order $e^{-\eta}$. Using the binomial expansion the first of the integrals may be written

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{(\eta+p)^n dp}{(e^p+1)(e^{-p}+1)} &= \sum_{r=0}^n a_r \eta^{n-r} \int_{-\infty}^{\infty} \frac{p^r dp}{(e^p+1)(e^{-p}+1)} \\ &= \sum_{r=0}^n a_r \eta^{n-r} \{1 + (-1)^r\} \int_0^{\infty} \frac{p^r dp}{(e^p+1)(e^{-p}+1)}, \end{aligned} \quad (3.10)$$

where

$$a_r = \frac{n!}{r!(n-r)!}.$$

The first integral in the sum in (3.10), corresponding to $r = 0$, is readily evaluated as $\frac{1}{2}$. The other integrals may be evaluated as follows (cf. the expression for $F_n(0)$, (3.6)):

$$\begin{aligned} \int_0^{\infty} \frac{p^r dp}{(e^p+1)(e^{-p}+1)} &= \left[-\frac{p^r}{e^p+1} \right]_0^{\infty} + \int_0^{\infty} \frac{r p^{r-1} dp}{e^p+1} \\ &= r(1-2^{1-r})(r-1)! \zeta(r). \end{aligned} \quad (3.11)$$

Combining (3.10), (3.11),

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{(\eta+p)^n dp}{(e^p+1)(e^{-p}+1)} &= \eta^n + \sum_{r=1}^n \frac{n!}{(n-r)!} \eta^{n-r} \{1 + (-1)^r\} (1-2^{1-r}) \zeta(r) \\ &= \eta^n \left\{ 1 + \sum_{r=1}^n 2n(n-1) \dots (n-2r+1) (1-2^{1-2r}) \zeta(2r) \eta^{-2r} \right\} \\ &= nS_{n-1}(\eta). \end{aligned} \quad (3.12)$$

The notation of the last line above is introduced for consistency later; the expression (3.12) provides the polynomial approximation for $I_n(\eta)$, and hence for $nF_{n-1}(\eta)$ (cf. (3.8)).

Consider now the function I_n with a negative argument

$$I_n(-\eta) = \int_0^{\infty} \frac{x^n dx}{(e^{x+\eta}+1)(e^{-x-\eta}+1)}. \quad (3.13)$$

By substituting $x+\eta = -p$ it may be shown from (3.13) that

$$I_n(-\eta) = (-)^n \int_{-\infty}^{-\eta} \frac{(\eta+p)^n dp}{(e^p+1)(e^{-p}+1)}. \quad (3.14)$$

Comparison of (3.9), (3.14) shows that, apart from the sign, (3.14) gives the term neglected in the derivation of the 'asymptotic series' for $I_n(\eta)$. Combining (3.9), (3.12), (3.14),

$$I_n(\eta) = nS_{n-1}(\eta) + (-)^{n+1} I_n(-\eta), \quad (3.15)$$

or

$$I_{n+1}(\eta) + (-)^{n+1} I_{n+1}(-\eta) = (n+1) S_n(\eta),$$

which with (3.8) gives

$$\left. \begin{aligned} F_n(\eta) + (-)^{n+1} F_n(-\eta) &= S_n(\eta), \\ \text{where } S_n(\eta) &= \frac{\eta^{n+1}}{n+1} \left\{ 1 + \sum_{r=1}^n 2(n+1)n \dots (n-2r+2) (1-2^{1-2r}) \zeta(2r) \eta^{-2r} \right\}. \end{aligned} \right\} \quad (3.16)$$

This is the required relation between $F_n(\eta)$ and $F_n(-\eta)$. The expression for $S_n(\eta)$ is equivalent to that for the asymptotic series given in M.S. p. 81, but, as mentioned previously, for positive integral values of n the series terminates after a finite number

of terms. It is perhaps necessary to emphasize here that (3.16) has been derived only for positive integral values of n ; in particular, the derivation does not apply to half-integral values. The expressions for $S_n(\eta)$ corresponding to $n = 1(1)4$ are shown in table 1 (§4). The values of the zeta functions required are given by

$$\zeta(2r) = \frac{2^{2r-1} \pi^{2r}}{(2r)!} B_r, \quad (3.17)$$

where B_r is a Bernoulli number (see Dwight 1945, pp. 9, 10). It is the first term of the expression for $S_n(\eta)$ which is usually referred to as the 'low-temperature approximation' to $F_n(\eta)$.

Once $F_n(-\eta)$ has been evaluated (see equation (3.5)), $F_n(\eta)$ may be computed to the same degree of absolute accuracy directly from (3.16). For small values of η , however, the number of terms required in the series (3.5) for $F_n(-\eta)$ is somewhat large, and to facilitate the use of the functions of particular physical interest $F_n(-\eta)$ has been calculated for $n = 1(1)4$, $\eta = 0.0(0.1)4.0$, and the values are tabulated in table 2 (§4).

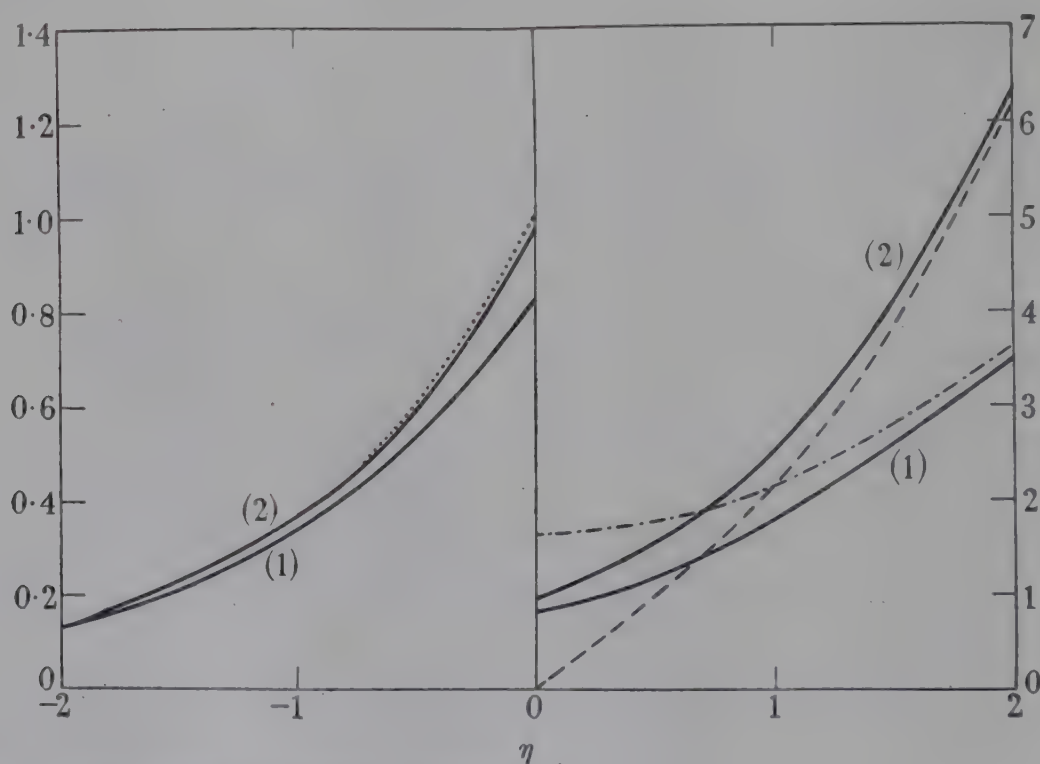


FIGURE 1. Fermi-Dirac and related functions

$$F_n(\eta) = F_n = \int_0^\infty \{x^n / (e^{x-\eta} + 1)\} dx.$$

Full curves: (1) F_1 ; (2) $(1/4!) F_4$. Dotted curve: e^η ; dot-dash curve: $S_1(\eta)$; broken curve: $(1/4!) S_4(\eta)$; $S_n(\eta)$, polynomial approximation to F_n , see table 1. The left-hand scale shows the ordinate scale for the curves for $\eta \leq 0$; the right-hand scale shows that for the curves for $\eta \geq 0$.

In illustration of the relation between the Fermi-Dirac functions and the approximations appropriate for $-\eta \gg 1$ and $\eta \gg 1$ the functions F_1 and $(1/4!) F_4$ are shown in figure 1, together with the corresponding approximations. For $\eta \leq 0$ the curve shown of e^η represents the high-temperature approximation to $(1/n!) F_n$ for all positive integral values of n (cf. (3.5)); the curves for $n > 4$ lie between the curve for

$(1/4!)F_4$ and that for e^η . For $\eta \geq 0$ the approximating curves correspond to the polynomial approximations (cf. table 1).

4. COMPUTATIONAL DETAILS AND NUMERICAL TABLES

The number of terms required in the series (3.5) to obtain any chosen accuracy increases rapidly as η approaches zero. For such values of η it is convenient to obtain the values of $F_n(\pm \eta)$ by a Taylor series, using values of $F_n(0)$, and of its derivatives, determined from (3.3), (3.6). From these two equations the r th derivative of $F_n(\eta)$ at $\eta = 0$ may be expressed as

$$F_n^{(r)}(0) = n!(1 - 2^{r-n})\zeta(n - r + 1). \quad (4.1)$$

Although the derivation of this equation is applicable only for $n > r$, comparison of the expressions obtained from it with those obtained by successive differentiations of the expression for $F_0, \ln(1 + e^\eta)$, indicates that it is valid for all values of r . The values of the zeta functions with positive integral arguments required in (4.1) are given by Davis (1933); those with zero and negative integral arguments may be obtained from the relations (Whittaker & Watson 1927, p. 268)

$$\left. \begin{aligned} \zeta(0) &= -\frac{1}{2}, & \zeta(-2m) &= 0, \\ \zeta(1-2m) &= (-)^m B_m/2m, \end{aligned} \right\} \quad (4.2)$$

where B_m is a Bernoulli number.

Values of $F_4(\pm \eta)$ have been calculated by the Taylor series for $\eta = 0.1, 0.2$. The values at 0.2 have been checked against those computed by direct summation of the series (3.5), and by means of (3.16). For $\eta \geq 0.2$ values of $F_4(-\eta)$ have been calculated from the series (3.5). The values of the exponentials required were taken from Newman (1883). Nine decimal places were used and sufficient terms were included so that the first term neglected was smaller than $\frac{1}{2}$ unit in the ninth decimal place. Similar methods were adopted for $F_3(\eta)$ and $F_2(\eta)$. In both cases the Taylor series were used for $\pm \eta = 0.1, 0.2, 0.3$. For F_3 the series (3.5) was used for $\eta \geq 0.3$; for F_2 the series was used for $\eta \geq 0.4$.

For $F_1(-\eta)$ the number of terms required in (3.5) is larger than for F_2, F_3, F_4 with corresponding values of η . Moreover, values of F_0 may be obtained directly from the expression $\ln(1 + e^\eta)$. In view of this it was considered most convenient, for $0.0 \leq \eta \leq 2.0$, to determine $F_1(\eta)$ from (3.4), using numerical integration. The function $\ln(1 + e^\eta)$ was tabulated at an interval of 0.1 in η , and values of the integral from 0 to η were obtained for $\eta = 0.1(0.1)2.0$ by summing the contributions found by step-by-step integration. The central-difference formula (cf. M.S. p. 74) was used for the integrations. For step-by-step integration this may be conveniently re-expressed

$$\text{as} \quad \frac{1}{w} \int_a^{a+w} f(x) dx = \frac{1}{2}(f_0 + f_1) - \frac{1}{24}(\delta_0^2 + \delta_1^2) + \frac{11}{1440}(\delta_0^4 + \delta_1^4) + \dots, \quad (4.3)$$

in which the notation follows that of M.S. The values of the exponentials required were taken from the W.P.A. tables (Lowan *et al.* 1939), and of the natural logarithms from the W.P.A. tables (Lowan *et al.* 1941). As a check on the numerical integrations the value of $F_1(2.0)$ calculated in this way was compared with the value of $F_1(2.0)$ calculated, by means of the relation (3.16), from the value of $F_1(-2.0)$ computed by

TABLE 1. POLYNOMIAL APPROXIMATION, $S_n(\eta)$, TO $F_n(\eta)$ FOR $\eta \gg 1$

$$S_n(\eta) = \frac{\eta^{n+1}}{n+1} \left\{ 1 + \sum_{r=1} 2(n+1)n \dots (n-2r+2) (1-2^{1-2r}) \zeta(2r) \eta^{-2r} \right\}.$$

$$(\pi^2 = 9.869\,6044; \pi^4 = 97.409\,0910.)$$

n	$S_n(\eta)$
1	$\frac{1}{2}\eta^2 + \frac{1}{6}\pi^2$
2	$\frac{1}{3}\eta^3 + \frac{1}{3}\pi^2\eta$
3	$\frac{1}{4}\eta^4 + \frac{1}{2}\pi^2\eta^2 + \frac{7}{60}\pi^4$
4	$\frac{1}{5}\eta^5 + \frac{2}{3}\pi^2\eta^3 + \frac{7}{15}\pi^4\eta$

TABLE 2. FERMI-DIRAC FUNCTIONS

$$F_n(\eta) = F_n = \int_0^\infty \frac{x^n dx}{e^{x-\eta} + 1}.$$

$$F_n(+|\eta|) = S_n(+|\eta|) + (-)^n F_n(-|\eta|); \text{ for } S_n(\eta) \text{ see table 1.}$$

$-\eta$	$\frac{1}{4!}F_4$	$\frac{1}{3!}F_3$	$\frac{1}{2!}F_2$	F_1
4.0	0.018 3052	0.018 2947	0.018 2739	0.018 2324
3.9	20 2291	20 2164	20 1910	20 1404
3.8	22 3552	22 3396	22 3086	22 2469
3.7	24 7045	24 6855	24 6477	24 5724
3.6	27 3005	27 2773	27 2311	27 1393
3.5	0.030 1690	0.030 1407	0.030 0844	0.029 9724
3.4	33 3386	33 3041	33 2354	33 0989
3.3	36 8409	36 7988	36 7150	36 5485
3.2	40 7106	40 6592	40 5570	40 3542
3.1	44 9862	44 9235	44 7988	44 5518
3.0	0.049 7101	0.049 6336	0.049 4817	0.049 1807
2.9	54 9293	54 8360	54 6508	54 2843
2.8	60 6954	60 5817	60 3560	59 9098
2.7	67 0656	66 9269	66 6519	66 1089
2.6	74 1028	73 9337	73 5987	72 9381
2.5	0.081 8767	0.081 6705	0.081 2626	0.080 4593
2.4	090 4638	090 2126	089 7159	088 7395
2.3	099 9488	099 6427	099 0382	097 8519
2.2	110 4249	110 0521	109 3166	107 8762
2.1	121 9952	121 5410	120 6467	118 8985
2.0	0.134 7728	0.134 2199	0.133 1327	0.131 0125
1.9	148 8828	148 2099	146 8889	144 3191
1.8	164 4629	163 6442	162 0399	158 9276
1.7	181 6647	180 6689	178 7217	174 9552
1.6	200 6550	199 4445	197 0825	192 5283
1.5	0.221 6178	0.220 1467	0.217 2834	0.211 7820
1.4	244 7550	242 9683	239 4997	232 8607
1.3	270 2891	268 1202	263 9215	255 9184
1.2	298 4644	295 8330	290 7548	281 1185
1.1	329 5494	326 3589	320 2224	308 6341
1.0	0.363 8391	0.359 9733	0.352 5649	0.338 6480
0.9	401 6570	396 9764	388 0416	371 3520
0.8	443 3582	437 6954	426 9316	406 9473
0.7	489 3318	482 4864	469 5344	445 6434
0.6	540 0041	531 7367	516 1710	487 6584
0.5	0.595 8417	0.585 8665	0.567 1842	0.533 2173
0.4	657 3552	645 3316	622 9403	582 5521
0.3	725 1023	710 6256	683 8284	635 9004
0.2	799 6923	782 2821	750 2622	693 5046
0.1	881 7900	860 8774	822 6794	755 6107
0.0	0.972 1198	0.947 0328	0.901 5427	0.822 4670

the series (3.5). The two values of $F_1(2.0)$ differed by less than 1 unit in the ninth decimal place. In order to obtain the functions in a form suitable for tabulation the values of $F_1(-\eta)$ for $\eta = 0.0(0.1)2.0$ were computed from the calculated values of $F_1(\eta)$ by means of the relation (3.16). For $2.0 \leq \eta \leq 4.0$ the values of $F_1(-\eta)$ were calculated directly from the series (3.5) as for F_2, F_3, F_4 .

In table 2 the functions F_4, F_3, F_2, F_1 are shown in the reduced form $(1/n!) F_n(-|\eta|)$ for $\eta = 0.0(0.1)4.0$. The values are given to seven decimal places and are rounded from values calculated using nine decimal places. The representation of the functions in the reduced form shown has the advantage that the values in any column are the derivatives of the function in the preceding column, since from (3.3)

$$\frac{d}{d\eta} \left\{ \frac{1}{n!} F_n(\eta) \right\} = \frac{1}{(n-1)!} F_{n-1}(\eta). \quad (4.4)$$

The first derivative of $F_1(\eta)$ is given by $F'_1 = F_0 = \ln(1 + e^\eta)$, and expressions for the further derivatives may be obtained directly from this. Consequently the methods of interpolation by means of a Taylor series discussed in M.S. (pp. 90 et seq.) may also be applied, *mutatis mutandis*, to the present table.

For negative values of η outside the range of the table values of F_n of comparable accuracy to those in the table may be obtained by using only three terms in the series (3.5). For positive values of η the values of F_n may be calculated from the corresponding values with negative η by means of the relation (3.16) and the expressions for S_n given in table 1. In this way $F_n(\eta)$ may be obtained for all possible values of η in the important cases of $n = 1(1)4$.

I am grateful to Professor E. C. Stoner, F.R.S., for his guidance and encouragement; also to Mr B. A. Lilley for helpful criticism; and to the Department of Scientific and Industrial Research for a grant.

REFERENCES

- Davis, H. T. (ed.) 1933 *Tables of the higher mathematical functions*, 1. Bloomington, Ind.: Principia Press.
- Dwight, H. B. 1945 *Tables of integrals and other mathematical data*. New York: Macmillan.
- Lowan, A. N. et al. 1939 *Tables of the exponential function*. New York: Work Projects Administration.
- Lowan, A. N. et al. 1941 *Table of natural logarithms*, 3, 4. New York: Work Projects Administration.
- McDougall, J. & Stoner, E. C. 1938 *Phil. Trans. A*, **237**, 67.
- Mestel, L. 1950 *Proc. Camb. Phil. Soc.* **46**, 331.
- Néel, L. 1938 *C.R. Acad. Sci., Paris*, **206**, 1471.
- Newman, F. W. 1883 *Trans. Camb. Phil. Soc.* **13**, 145.
- Nordheim, L. 1934 Müller Pouillet's *Lehrbuch der Physik*, 4/4, 271. Brunswick: Vieweg.
- Putley, E. H. 1949 *Proc. Phys. Soc. A*, **62**, 284.
- Rhodes, P. 1950 *Proc. Roy. Soc. A*, **202**, 466.
- Sommerfeld, A. & Bethe, H. 1933 *Handb. Phys.* **24/2**, 333.
- Stoner, E. C. 1938 *Proc. Roy. Soc. A*, **165**, 372.
- Stoner, E. C. 1939a *Phil. Mag* [7] **28**, 257.
- Stoner, E. C. 1939b *Proc. Roy. Soc. A*, **169**, 339.
- Whittaker, E. T. & Watson, G. N. 1927 *A course of modern analysis*, 4th ed. Cambridge University Press.
- Wilson, A. H. 1936 *The theory of metals*. Cambridge University Press.
- Wilson, A. H. 1937 *Proc. Camb. Phil. Soc.* **33**, 371.

Theory of light absorption and non-radiative transitions in F -centres

BY KUN HUANG AND AVRIL RHYS

Department of Theoretical Physics, University of Liverpool

(Communicated by N. F. Mott, F.R.S.—Received 1 August 1950)

A quantitative theory for the shapes of the absorption bands of F -centres is given on the basis of the Franck-Condon principle. Underlying the treatment are two simplifying assumptions: namely, (a) that the lattice can be approximately treated as a dielectric continuum; (b) that in obtaining the vibrational wave functions for the lattice, the effect of the F -centre can be considered as that of a static charge distribution. Under these assumptions, it is shown that the absorption constant as a function of frequency and temperature can be expressed in terms of the Bessel functions with imaginary arguments. The theoretical curves for the absorption constant compare very favourably with the experimental curves for all temperatures.

Also considered in the paper are the probabilities of non-radiative transitions, which are important in connexion with the photo-conductivity observed following light absorption by F -centres. The treatment given differs from the qualitative considerations hitherto in one important aspect, namely, the strength of the coupling between the electron and the lattice is taken into account. The adiabatic wave functions for the F -centre electron required for the discussion are obtained by perturbation methods. The probability for an excited F -centre to return to its ground state by non-radiative transitions is shown to be negligible; similar transitions to the conduction band are, however, important if the excited state is separated from the conduction band by not much more than 0.1 eV. The temperature dependence of such transitions is complicated, but, for a wide range of temperatures, resembles $e^{-W/kT}$. Tentative estimates show that the result is consistent with the observed steep drop of the photo-conductive current with temperature.

1. INTRODUCTION

It has long been recognized that the considerable widths of the characteristic absorption curves of F -centres (Pohl 1937) are caused by the coupling of the electronic motion in the F -centre to the ionic lattice. Owing to the coupling, while an F -centre electron is making a transition, a number of lattice vibration quanta can be created or annihilated. Thus for the same electronic transition a certain latitude in the absorption frequency results, corresponding to a greater or lesser number of vibrational quanta being annihilated or created during the absorption act.

Although the basic mechanism underlying the absorption is clear, no quantitative theory of the absorption curves has hitherto been given. The usual interpretation of the optical properties of F -centres is based upon a schematic application of the Franck-Condon principle, whereby a single parameter is used to specify schematically the lattice configuration (von Hippel 1936; Seitz 1939; Mott & Gurney 1948). Apart from the natural limitations of the procedure, the principle is often indifferently interpreted so that important features could well be missed (no account, for instance, is usually taken of the detailed nature of the vibrational wave functions, which are, however, well known to be important for an understanding of the intensity distributions of the absorption spectra of molecules (cf. Finkelburg 1938)). In a recent paper Muto (1949) has attempted to give a quantitative treatment of the problem.

He has, however, proceeded no further than giving the general expressions for the transition probabilities. His results remain so formal that no actual comparison with experiments can be made. The only concrete conclusion he has drawn from the theory, namely, that the maximum of absorption corresponds to no change in the vibrational quanta, is in fact, as we shall see, incorrect.

General references in this connexion have not infrequently been made to the works of Frenkel (1936) and Peierls (1932), which are mainly concerned with a different problem, namely, the absorption by pure crystals. An important difference, however, exists between an impurity centre such as an F -centre and an exciton which is responsible for absorption in pure crystals. Owing to the high mobility of the latter and its internal motion shared by both the positive hole and the electron, an exciton cannot polarize the lattice to an appreciable extent. An impurity centre, on the other hand, is practically immobile; purely as a static charge distribution, it can polarize strongly the lattice in its neighbourhood. It is our belief that such a polarizing effect of the F -centre is mainly responsible for the observed course of absorption.

In this paper we shall give a quantitative theory of the absorption constant on the basis of the Franck-Condon principle, the effect of the F -centres on the lattice being considered as that of static charge distributions. It is found that if the lattice is treated approximately as a dielectric continuum, the result contains but few constants that cannot be reliably calculated on purely theoretical ground. Once the values of these constants are assigned, the absorption constant can be calculated for all temperatures and frequencies. The theoretical absorption curves are found to be in good agreement with the experimental curves (Pohl 1937).

The light absorption by F -centres is known experimentally to be associated with photoconductivity (Pohl 1937). But below a certain temperature in the liquid-air region, the photo-current is observed to drop very steeply with temperature. The problem has been discussed by Mott and Gurney (Gurney & Mott 1938; Mott 1938), who suggest that an F -centre electron raised to the excited state by light absorption can be thrown into the conduction band by lattice vibrations before it has the opportunity to return to the ground state. These authors have tentatively assumed that such non-radiative transitions vary with temperature as $\exp(-W/kT)$. By an extension of the above theory, the transition probabilities can be calculated. It will be seen that for a wide range of temperatures the probabilities follow closely the exponential form. The results in general are in agreement with the experiments on photoconductivity.

2. GENERAL EXPRESSION FOR THE ABSORPTION CONSTANT

Consider generally an absorption centre in vacuum, upon which falls a nearly homogeneous beam of radiation of unit cross-section, with $\rho(\nu)$ practically uniform in the frequency range $\nu \rightarrow \nu + \Delta\nu$ and vanishing at other frequencies ($\rho(\nu)$ denotes as usual the energy density per unit frequency range). The rate of energy absorption is then given by

$$\frac{8\pi^3}{3h} \nu \rho(\nu) \sum_f^{\nu < \nu_f < \nu + \Delta\nu} |\langle i | \mathbf{M} | f \rangle|^2, \quad (2.1)$$

where $\langle i | \mathbf{M} | f \rangle$ is the matrix element of the electric moment between the initial state i and a final state f of the absorption centre; as indicated, the summation is over such final states that the transition frequency ν_{fi} lies within the relevant range. However, if the absorption is imbedded in a refractive medium, $\rho(\nu)$ should be identified with $(\mathbf{E}^2/4\pi)$ rather than the energy density. In a medium with refractive index n , the intensity of radiation is given by

$$I = \frac{nc\mathbf{E}^2}{4\pi}.$$

(2.1) can thus be written as

$$\frac{8\pi^3\nu I}{3hnc} \left\{ \frac{1}{\Delta\nu} \sum_f^{\nu < \nu_{fi} < \nu + \Delta\nu} |\langle i | \mathbf{M} | f \rangle|^2 \right\}. \quad (2.2)$$

Let us consider a single F -centre in the lattice. A state of the whole system is specified by two quantum numbers: μ for the electronic states of the F -centre and n for the vibrational states of the lattice. The energy of a state will thus be written as $E_{\mu n}$. Following the usual approximation in the quantum-mechanical application of the Franck-Condon principle (Condon & Morse 1929), we write

$$\langle i | \mathbf{M} | f \rangle \rightarrow \langle \mu' | e\mathbf{x} | \mu'' \rangle \int X_{\mu'n'}^*(X) X_{\mu''n''}(X) dX, \quad (2.3)$$

where $\langle \mu' | e\mathbf{x} | \mu'' \rangle$ is the matrix element of the electric moment of the F -centre electron between the two electronic states μ', μ'' of the F -centre which are assumed responsible for the observed absorption. X in the formula stands symbolically for the nuclear co-ordinates specifying the lattice configurations, and $X_{\mu'n'}$, $X_{\mu''n''}$ are the vibrational wave functions of the lattice, when the F -centre is respectively in its initial and final states. If there are a number of F -centres present, their energy absorptions can clearly be considered additively. Let C be the number of centres per unit volume and x the direction of the incident beam. On combining (2.3) and (2.2), one obtains the rate of absorption in a layer of thickness dx in the following form:

$$-dI = \frac{8\pi^3C}{3hnc} |\langle \mu' | e\mathbf{x} | \mu'' \rangle|^2 \nu F(\nu) I dx, \quad (2.4)$$

where $F(\nu)$ is the function of frequency defined by

$$F(\nu) = \frac{1}{\Delta\nu} \sum_{n''}^{(\nu - \nu + \Delta\nu)} \left| \int X_{\mu'n'}^*(X) X_{\mu''n''}(X) dX \right|^2. \quad (2.5)$$

n'' is summed over final vibrational states with frequencies $h^{-1}(E_{\mu''n''} - E_{\mu'n'})$ within a range $\Delta\nu$ about ν . The absorption constant is thus given by

$$k(\nu) = \frac{8\pi^3C}{3hnc} |\langle \mu' | e\mathbf{x} | \mu'' \rangle|^2 \nu F(\nu). \quad (2.6)$$

The absorption curves depend essentially on $F(\nu)$, which, we notice, depend in turn on the overlap integrals between vibrational wave functions of the lattice for two different states of the F -centre.

3. CONTINUUM APPROXIMATION FOR THE LATTICE AND THE VIBRATIONAL WAVE FUNCTIONS

We shall assume that the effect of the F -centre on the lattice is due purely to the Coulomb field $\mathcal{E}_\mu$ (or $\mathcal{E}_\mu^*$) of the average charge distribution in the F -centre; the suffix to the field indicates the state of the F -centre. To avoid introducing many unknown parameters into the theory, we shall consider the lattice essentially as a dielectric continuum. The behaviour of such a continuum lattice is completely described by a pair of equations given by Huang, (1950). In the continuum approximation, apart from purely elastic deformations which are irrelevant here, the lattice configuration is specified by giving over all positions $\mathbf{r}$ in the lattice the displacement vector $\mathbf{u}(\mathbf{r})$ of the positive ions relative to the negative ions. Using Huang's equations and following the method exemplified in his paper, it is not difficult to show that the total energy of the lattice (potential energy counted from the configuration $\mathbf{u}(\mathbf{r}) \equiv 0$) in the presence of an external electric field $\mathcal{E}_\mu^*(\mathbf{r})$ is given by

$$\frac{\bar{M}}{2} \int \{ \dot{\mathbf{u}}_l^2 + \omega_l^2 \mathbf{u}_l^2 + \dot{\mathbf{u}}_t^2 + \omega_0^2 \mathbf{u}_t^2 \} \frac{dv}{v_a} - \omega_l \left[\frac{\bar{M}}{4\pi v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \right]^{\frac{1}{2}} \int \mathbf{u}_l \cdot \mathcal{E}_\mu^* dv, \quad (3.1)$$

where $\mathbf{u}_l(\mathbf{r})$ and $\mathbf{u}_t(\mathbf{r})$ are respectively the irrotational and solenoidal parts of $\mathbf{u}(\mathbf{r})$, and

$$\left. \begin{aligned} \epsilon_0 &= \text{static dielectric constant,} \\ \epsilon_\infty &= \text{high frequency dielectric constant,} \\ \omega_0 &= \text{infra-red dispersion frequency (circular),} \\ \bar{M} &= \text{reduced mass of the ions in the cell } M_+ M_- / (M_+ + M_-), \\ v_a &= \text{volume of unit cell,} \\ \omega_l &= (\epsilon_0 / \epsilon_\infty)^{\frac{1}{2}} \omega_0. \end{aligned} \right\} \quad (3.2)$$

We observe that $\mathbf{u}_l(\mathbf{r})$ is not coupled to the F -centre field; this part of $\mathbf{u}(\mathbf{r})$ gives rise to independent modes of vibrations (transverse waves) irrelevant for our discussion, and will henceforth be ignored.

The energy expression (3.1) can be normalized to a finite volume by imposing the Born-Kármán boundary conditions on a region containing $L \times L \times L = N$ cells and restricting the volume integrations in (3.1) to the same. Thus it is convenient to write

$$\sqrt{N} \mathbf{u}_l(\mathbf{r}) = \left(\frac{1}{\bar{M}} \right)^{\frac{1}{2}} \sum_{\mathbf{y}} Q(\mathbf{y}) (\mathbf{y} / |\mathbf{y}|) \exp [2\pi i \mathbf{y} \cdot \mathbf{r}], \quad (3.3)$$

where we have resolved $\mathbf{u}_l(\mathbf{r})$ into longitudinal waves; this is possible because of the irrotational character of $\mathbf{u}_l(\mathbf{r})$. The condition of reality permits us to write

$$Q(\mathbf{y}) = \frac{1}{\sqrt{2}} (q_{y1} + i q_{y2}) = -Q^*(-\mathbf{y}). \quad (3.4)$$

The Born-Kármán conditions can be expressed by restricting the values of $\mathbf{y}$ as follows:

$$\mathbf{y} = L^{-1} (h^1 \mathbf{b}_1 + h^2 \mathbf{b}_2 + h^3 \mathbf{b}_3) \quad |h^i| = \text{integers} < \frac{1}{2} L, \quad (3.5)$$

where $\mathbf{b}^1, \mathbf{b}^2, \mathbf{b}^3$ are the reciprocal basic vectors of the lattice (only in a genuine continuum does $|h^i| > \frac{1}{2} L$ lead to further distinct modes of displacements). On

substituting (3.3) in (3.1), ignoring $\mathbf{u}_l(\mathbf{r})$, and carrying out the volume integration over the region of N cells, one obtains

$$H = \frac{1}{2} \sum_{\lambda=1,2} \sum_{\mathbf{y}} (\dot{q}_{\mathbf{y}\lambda}^2 + \omega_l^2 q_{\mathbf{y}\lambda}^2) - \frac{1}{\sqrt{N}} \sum_{\mathbf{y}\lambda} A'_{\mathbf{y}\lambda} q_{\mathbf{y}\lambda}, \quad (3.6)$$

where

$$\left. \begin{aligned} A'_{\mathbf{y}1} &= \omega_l \left[\frac{1}{2\pi v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \right]^{\frac{1}{2}} \int (\mathbf{y} / |\mathbf{y}|) \cdot \mathcal{E}_{\mu'}(\mathbf{r}) \cos 2\pi \mathbf{y} \cdot \mathbf{r} dv, \\ A'_{\mathbf{y}2} &= -\omega_l \left[\frac{1}{2\pi v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \right]^{\frac{1}{2}} \int (\mathbf{y} / |\mathbf{y}|) \cdot \mathcal{E}_{\mu'}(\mathbf{r}) \sin 2\pi \mathbf{y} \cdot \mathbf{r} dv. \end{aligned} \right\} \quad (3.7)$$

Owing to the combination of terms arising from $Q(\mathbf{y})$ and $Q(-\mathbf{y})$, the summation over $\mathbf{y}$ now extends over the values of $\mathbf{y}$ lying on one side of an arbitrary plane through the origin of the $\mathbf{y}$ -space; in this way only one of a pair $\mathbf{y}$, $-\mathbf{y}$ is counted. For simplicity, we introduce a single index j for $\mathbf{y}\lambda$; (3.6), for instance, becomes

$$H = \frac{1}{2} \sum_j (\dot{q}_j^2 + \omega_l^2 q_j^2) - \frac{1}{\sqrt{N}} \sum_j A'_j q_j. \quad (3.8)$$

In the absence of the F -centre, the linear term in (3.8) drops out. The Hamiltonian describes then a system of independent oscillators, all having the same frequency ω_l (circular). q_j are thus the normal co-ordinates of the perfect lattice; the corresponding modes have the form of longitudinal waves.

We notice that the linear terms in (3.8) represent the coupling between the F -centre and the lattice and are proportional to $1/\sqrt{N}$. It is thus mathematically important to have normalized the lattice to a finite volume. In the result, we can, of course, let $N \rightarrow \infty$, as we are not concerned with surface effects.

When the F -centre is present and in the state μ' , the following modified normal co-ordinates can be introduced:

$$q'_j = q_j - \frac{1}{\sqrt{N}} \frac{A'_j}{\omega_l^2}. \quad (3.9)$$

For on substituting (3.9) in (3.8), one finds

$$H = \frac{1}{2} \sum_j (\dot{q}'_j{}^2 + \omega_l^2 q'_j{}^2) - \frac{1}{2N} \sum_j \frac{A'_j{}^2}{\omega_l^2}. \quad (3.10)$$

The Hamiltonian is again that of a system of simple oscillators, apart from the last term, which represents the equilibrium potential energy between the F -centre and the lattice. It is usual to include the latter as part of the electronic energy of the state μ' of the F -centre. Thus the first part alone in (3.10) will be considered as the lattice vibration energy.

$\dot{q}'_j$ is classically the canonical conjugate momentum of q'_j ; thus on going over to quantum mechanics, the vibrational Hamiltonian of the lattice becomes

$$H_{\text{vib.}} = \frac{1}{2} \sum_j \left(-\hbar^2 \frac{\partial^2}{\partial q'^2_j} + \omega_l^2 q'^2_j \right). \quad (3.11)$$

The corresponding wave functions are the products of the oscillator wave functions

$$\prod_j X_{n'_j}(q'_j), \quad (3.12)$$

where n'_j is the quantum number for the oscillator j ; the corresponding vibrational energy is the sum

$$\hbar \omega_l \sum_j (n'_j + \frac{1}{2}). \quad (3.13)$$

4. CALCULATION OF $F(\nu)$

Consider the overlap integrals required for constructing the function $F(\nu)$:

$$\int \mathbf{X}_{\mu'n'}^* \mathbf{X}_{\mu'n} d\mathbf{X} = \prod_j \int \mathbf{X}_{n'_j}(q'_j) \mathbf{X}_{n''_j}(q''_j) dq_j, \quad (4.1)$$

where the vibrational wave functions (3.12) have been substituted and q'_j, q''_j refer respectively to the modified normal co-ordinates of the lattice when the F -centre is in state μ' and state μ'' .

Remembering that

$$q'_j = q_j - \frac{1}{\sqrt{N}} \frac{A'_j}{\omega_l^2}, \quad q''_j = q_j - \frac{1}{\sqrt{N}} \frac{A''_j}{\omega_l^2},$$

we expand the oscillator wave functions in (4.1) about q_j , thus

$$\begin{aligned} & \int \mathbf{X}_{n'_j}(q'_j) \mathbf{X}_{n''_j}(q''_j) dq_j \\ &= \int \left\{ \left[\mathbf{X}_{n'_j}(q_j) - \frac{1}{\sqrt{N}} \frac{A'_j}{\omega_l^2} \mathbf{X}'_{n'_j}(q_j) + \frac{1}{2} \left(\frac{1}{\sqrt{N}} \frac{A'_j}{\omega_l^2} \right)^2 \mathbf{X}''_{n'_j}(q_j) + \dots \right] \right. \\ & \quad \times \left. \left[\mathbf{X}_{n''_j}(q_j) - \frac{1}{\sqrt{N}} \frac{A''_j}{\omega_l^2} \mathbf{X}'_{n''_j}(q_j) + \frac{1}{2} \left(\frac{1}{\sqrt{N}} \frac{A''_j}{\omega_l^2} \right)^2 \mathbf{X}''_{n''_j}(q_j) + \dots \right] \right\} dq_j. \end{aligned} \quad (4.2)$$

When the multiplication of the two series is carried out, one observes that the terms of order $(1/\sqrt{N})^s$ are associated with integrals of the form

$$\int \mathbf{X}_{n'_j}^{(l)} \mathbf{X}_{n''_j}^{(s-l)} dq_j$$

(superscripts indicate the order of derivative), all of which can, by integration by parts, be expressed in terms of

$$\int \mathbf{X}_{n'_j} \mathbf{X}_{n''_j}^{(s)} dq_j.$$

This integral vanishes, whenever the difference between n'_j and n''_j is numerically greater than s . Thus if $|n'_j - n''_j| = s$, the leading term of the integral (4.2) is of the order $(1/\sqrt{N})^s$. As we shall later see, it follows from this fact that all transitions in which any of the vibrational quantum numbers changes by more than one can be ignored.

For the case $\Delta n_j = n''_j - n'_j = \pm 1$, one finds on multiplying out the series in (4.2) that

$$\begin{aligned} & \int \mathbf{X}_{n'_j}(q'_j) \mathbf{X}_{n'_j \pm 1}(q''_j) dq_j \\ &= -\frac{1}{\sqrt{N} \omega_l^2} \left[A'_j \int \mathbf{X}'_{n'_j}(q_j) \mathbf{X}_{n'_j \pm 1}(q_j) dq_j + A''_j \int \mathbf{X}_{n'_j}(q_j) \mathbf{X}'_{n'_j \pm 1}(q_j) dq_j \right] + O \frac{1}{N}, \quad \text{etc.} \\ &= \frac{A''_j - A'_j}{\sqrt{N}} \left(\frac{\hbar}{2\omega_l^3} \right)^{\frac{1}{2}} \left\{ \sqrt{(n'_j + 1)} - \sqrt{n'_j} \right\} + O \frac{1}{N}, \quad \text{etc.} \end{aligned} \quad (4.3)$$

The leading term is the only one that we shall require.

Let $E_{\mu''\mu'}$ be the difference of electronic energies between the states μ'' and μ' . The transition frequency for a particular transition $\mu'n' \rightarrow \mu''n''$ is given by (cf. (3.13))

$$h\nu = E_{\mu''\mu'} + \hbar\omega_l \sum_j \Delta n_j, \quad \text{where} \quad \Delta n_j = n''_j - n'_j. \quad (4.4)$$

In our simplified model, thus ν depends only on the change in the total number of vibrational quanta, and, moreover, ν does not cover a continuum when n'' goes through all possible final vibrational states, but only takes on the discrete values

$$\nu = \frac{E_{\mu''\mu'}}{h} + p \left(\frac{\omega_l}{2\pi} \right), \quad (4.5)$$

where p is an integer, equal to $\sum_j \Delta n_j$. The function $F(\nu)$ as defined in (2.6) can obviously be rewritten in this case as

$$F(\nu) = \frac{2\pi}{\omega_l} \left\{ \sum_{n''} \left| \int X_{\mu''n''}^* X_{\mu'n'} dX \right|^2 \right\}_{\sum_j \Delta n_j = p}. \quad (4.6)$$

As indicated, the summation over n'' is restricted by the condition $\sum_j \Delta n_j = p$, where p and the absorption frequency ν are related by (4.5).

Consider first transitions in which no $|\Delta n_j|$ is greater than one. Transitions of this type that contribute to (4.6) are made up of s oscillators going down by one quantum and $s + p$ oscillators going up by one quantum. Let us consider a particular transition of this type, where the oscillators going up and down are respectively $l', l'', \dots, l^{(s+p)}$ and $k', k'', \dots, k^{(s)}$. The square of the overlap integral (4.1) for this transition can be written as

$$\begin{aligned} \left| \int X_{\mu''n''}^* X_{\mu'n'} dX \right|^2 &= \left\{ \prod_j \left| \int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j \right|^2 \right\} \\ &\times \left\{ \prod_{l'} \frac{\left| \int X_{n'_l}(q'_l) X_{n'_{l+1}}(q''_l) dq_l \right|^2}{\left| \int X_{n'_l}(q'_l) X_{n'_l}(q''_l) dq_l \right|^2} \right\} \left\{ \prod_k \frac{\left| \int X_{n'_k}(q'_k) X_{n'_{k-1}}(q''_k) dq_k \right|^2}{\left| \int X_{n'_k}(q'_k) X_{n'_k}(q''_k) dq_k \right|^2} \right\}, \end{aligned} \quad (4.7)$$

where the products range respectively over all the oscillators, and the oscillators $l', l'', \dots$ and $k', k'', \dots$. For given p and s , there are a large number of distinct transitions, corresponding to different choices of the oscillators $l', l'', \dots$ and $k', k'', \dots$. Their total contribution can be obtained by summing (4.7) over all the indices $l', l'', \dots$ and $k', k'', \dots$ and dividing afterwards by $s!(s+p)!$; the latter division takes account of the repetition of each distinct transition implied in the straightforward summation. The summation obtained in this way formally also includes terms corresponding to some of the indices $l', l'', \dots, k', k'', \dots$ being equal, which is meaningless. But it will become evident that these terms can make no finite difference to the result. Thus we find that the total contribution to $F(\nu)$ by transitions in which $s + p$ oscillators go up by one quantum and s oscillators go down by one quantum can be written as

$$\begin{aligned} \frac{2\pi}{\omega_l} \frac{1}{s!(s+p)!} &\left\{ \prod_j \left| \int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j \right|^2 \right\} \\ &\times \left\{ \sum_{l'} \frac{\left| \int X_{n'_l}(q'_l) X_{n'_{l+1}}(q''_l) dq_l \right|^2}{\left| \int X_{n'_l}(q'_l) X_{n'_l}(q''_l) dq_l \right|^2} \right\}^{s+p} \\ &\left\{ \sum_k \frac{\left| \int X_{n'_k}(q'_k) X_{n'_{k-1}}(q''_k) dq_k \right|^2}{\left| \int X_{n'_k}(q'_k) X_{n'_k}(q''_k) dq_k \right|^2} \right\}^s \end{aligned} \quad (4.8)$$

The summations over l and k are both extended over all the lattice oscillators; if we return to the indices y, λ for the lattice oscillators the summations can be replaced by the integrations

$$\sum_{y\lambda} \rightarrow \sum_{\lambda} \int dh^1 dh^2 dh^3 = N v_a \sum_{\lambda} \int dy. \quad (4.9)$$

The leading term of the corresponding integrand

$$\frac{\left| \int X_{n'_l}(q'_l) X_{n'_l \pm 1}(q''_l) dq_l \right|^2}{\left| \int X_{n'_l}(q'_l) X_{n_l}(q''_l) dq_l \right|^2},$$

is, according to (4.2), (4.3), given by

$$\frac{(A''_l - A'_l)^2}{N} \left(\frac{\hbar}{2\omega_l^3} \right) \frac{n'_l + 1}{n'_l}. \quad (4.10)$$

N in the denominator cancels the factor outside the y -integral (4.9) and leads thus to a finite contribution; besides the leading term the rest of the expansion contributes only by terms of order $(1/\sqrt{N})$ or higher, and may thus be ignored. Therefore (4.8) can be written as

$$\frac{2\pi}{\omega_l} \frac{1}{s!(s+p)!} \prod_j \left| \int X_{n'_j}(q'_j) X_{n_j}(q''_j) dq_j \right|^2 S_+^{s+p} S_-^s, \quad (4.11)$$

where $S_{\pm}$ are the following integrals:

$$S_{\pm} = \frac{\hbar v_a}{2\omega_l^3} \sum_{\lambda} \int (A''_{y\lambda} - A'_{y\lambda})^2 \frac{n'_{y\lambda} + 1}{n'_{y\lambda}} dy. \quad (4.12)$$

As regards transitions involving certain oscillators changing by more than one quantum, let us consider, for instance, transitions in which certain s' oscillators go up by two quanta. Corresponding to (4.8), we would obtain an expression in which appears the factor

$$\left\{ N \sum_{\lambda} \int dy \frac{\left| X_{n'_{y\lambda}}(q'_{y\lambda}) X_{n'_{y\lambda}+2}(q''_{y\lambda}) dq_{y\lambda} \right|^2}{\left| X_{n'_{y\lambda}}(q'_{y\lambda}) X_{n'_{y\lambda}}(q''_{y\lambda}) dq_{y\lambda} \right|^2} \right\}.$$

It follows from our earlier explanations that the integrand of the y -integration will be of the order $(1/N)^2$; thus the integral will be of order $(1/N)$. Results of even higher orders will be obtained, if one considers some oscillators changing by more than two quanta. Hence no finite contributions arise from such transitions.

All relevant contributions to $F(\nu)$ are thus included, if we sum (4.11) over various values of s :

$$F(\nu) = \frac{2\pi}{\omega_l} \prod_j \left| \int X_{n'_j}(q'_j) X_{n_j}(q''_j) dq_j \right|^2 \sum_{s=0}^{\infty} \frac{S_+^{s+p} S_-^s}{s!(s+p)!}. \quad (4.13)$$

where, we remember, ν and p are related by (4.5). The sum can be expressed in terms of the Bessel function with imaginary argument of order p (Whittaker & Watson 1946), for

$$\sum_{s=0}^{\infty} \frac{S_+^{s+p} S_-^s}{s!(s+p)!} = \left(\frac{S_+}{S_-} \right)^{\frac{1}{2}p} \sum_{s=0}^{\infty} \frac{(S_+ S_-)^{s+\frac{1}{2}p}}{s!(s+p)!} = \left(\frac{S_+}{S_-} \right)^{\frac{1}{2}p} I_p(2\sqrt{(S_+ S_-)}). \quad (4.14)$$

Hence (4.13) may be written alternatively as

$$F(\nu) = \frac{2\pi}{\omega_l} \left\{ \prod_j \left| \int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j \right|^2 \right\} \left(\frac{S_+}{S_-} \right)^{\frac{1}{2}p} I_p(2\sqrt{(S_+S_-)}). \quad (4.15)$$

We have tacitly assumed that $p \geq 0$, i.e. the number of oscillators going up equals or exceeds the number going down. For $p < 0$, i.e. the number of oscillators going down exceeds the number going up by $|p| = -p$, the previous arguments apply, if we interchange the roles of the two sets of oscillators; the net effect on the result is equivalent to interchanging S_+ and S_- . Thus for $p < 0$, we would obtain

$$F(\nu) = \frac{2\pi}{\omega_l} \left\{ \prod_j \left| \int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j \right|^2 \right\} \left(\frac{S_-}{S_+} \right)^{-\frac{1}{2}p} I_{-p}(2\sqrt{(S_+S_-)})$$

However as $I_{-p} = I_p$ (Whittaker & Watson 1946), this formula is identical with (4.15) which holds thus equally for positive and negative values of p .

The value of the product

$$\prod_j \left| \int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j \right|^2 \quad (4.16)$$

remains to be evaluated. With the help of (4.2), one can express

$$\int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j$$

in a series in $(1/N)$:

$$1 - \frac{1}{2N} (A''_j - A'_j)^2 \left(\frac{\hbar}{2\omega_l^3} \right) (2n'_j + 1) + O \frac{1}{N^2}, \quad \text{etc.}$$

On squaring the series, one finds

$$\left| \int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j \right|^2 = 1 - \frac{1}{N} (A''_j - A'_j)^2 \left(\frac{\hbar}{2\omega_l^3} \right) (2n'_j + 1) + O \frac{1}{N^2}, \quad \text{etc.} \quad (4.17)$$

When the factors in (4.16) are expanded in this way, it is readily seen on multiplying out the factors that terms of order $(1/N)^2$ or higher in the factors can be ignored without making any finite difference to the result. Thus retaining in each factor only the two terms, which are explicitly given in (4.17), one finds that the considerations required for working out the product in (4.16) reduce to essentially the same as that used earlier in this section. It is easily verified that the result is

$$\prod_j \left| \int X_{n'_j}(q'_j) X_{n''_j}(q''_j) dq_j \right|^2 = \sum_{t=0}^{\infty} \frac{(-S_+ - S_-)^t}{t!} = \exp[-(S_+ + S_-)]. \quad (4.18)$$

When (4.18) is substituted, (4.15) becomes

$$F(\nu) = \frac{2\pi}{\omega_l} \left(\frac{S_+}{S_-} \right)^{\frac{1}{2}p} \exp[-(S_+ + S_-)] I_p(2\sqrt{(S_+S_-)}). \quad (4.19)$$

When we average the initial state over the thermal distribution, the quantities S_+ , S_- defined in (4.12) become

$$S_{\pm} = S \left\{ \frac{\bar{n}(T) + 1}{\bar{n}(T)} \right\}, \quad (4.20)$$

where S is the constant integral

$$S = \frac{\hbar\nu_a}{2\omega_l^3} \sum_{\lambda} \int (A''_{\mathbf{y}\lambda} - A'_{\mathbf{y}\lambda})^2 d\mathbf{y}, \quad (4.21)$$

and $\bar{n}(T)$ is the thermal average of the vibrational quantum number which is the same for all oscillators in our case:

$$\bar{n}(T) = \{\exp[\hbar\omega_l/kT] - 1\}^{-1}. \quad (4.22)$$

When the expression (4.19) for $F(\nu)$ is substituted in (2.6) and $S_{\pm}$ are expressed in terms of S and $\bar{n}(T)$ by (4.20), we obtain eventually the following expression for the absorption constant as a function of frequency and temperature:

$$k(\nu) = A\nu \exp[-S(2\bar{n} + 1)] \left(\frac{\bar{n} + 1}{\bar{n}}\right)^{\frac{1}{2}p} I_p(2S\sqrt{[\bar{n}(\bar{n} + 1)]}). \quad (4.23)$$

$$A = \frac{16\pi^4 C}{3\hbar n c \omega_l} |\langle \mu' | e\mathbf{x} | \mu'' \rangle|^2.$$

5. COMPARISON WITH EXPERIMENTAL RESULTS

We can disregard for a moment the absolute frequency scale, if we divide the absorption constant k by ν :

$$k/\nu = A \exp[-S(2\bar{n} + 1)] \left(\frac{\bar{n} + 1}{\bar{n}}\right)^{\frac{1}{2}p} I_p(2S\sqrt{[\bar{n}(\bar{n} + 1)]}), \quad (5.1)$$

and plot k/ν against $p\hbar\omega_l$. For the large arguments of I_p , with which we shall be concerned ($S \sim 20$), no tabulated values or satisfactory asymptotic formulae for the function are available for the whole desired range for p . The values of the function have thus to be obtained by laborious computations; and we have confined our comparisons to the often-quoted absorption curves for KBr due to Pohl (1937).

As Pohl's curves are plotted on an arbitrary scale, for the comparison the constant A in (5.1) serves merely as an arbitrary scale factor. It is found that with the value of S suitably chosen ($S = 22.4$) the theoretical curves for k/ν can be fitted on to Pohl's curves (reduced by ν) with excellent agreement for all temperatures. The arbitrary scale constant A is of course kept at the same value for the theoretical curves corresponding to different temperatures.

I_p , as a function of its order (for fixed argument), has a maximum at $p = 0$. The maximum is shifted towards positive values of p , after the multiplication by $[(\bar{n}(T) + 1)/\bar{n}(T)]^{\frac{1}{2}p}$. When the value of $\bar{n}(T)$ is raised by temperature, the latter factor approaches nearer to unity; however, owing to the raised value of the argument $2S[\bar{n}(\bar{n} + 1)]^{\frac{1}{2}}$, I_p itself has a flatter maximum, and one finds in fact that the maximum of the curve $k/\nu \sim p\hbar\omega_l$ occurs practically at the same value of p ($\simeq 22$) for all the temperatures considered. Since the maxima of Pohl's curves shift with temperature, it becomes evident that no choice of the difference $E_{\mu',\mu''}$ of electronic energies (cf. (4.5)) will bring agreement between the absolute positions of the theoretical and experimental curves on the frequency scale. Owing to the otherwise satisfactory agreement between the two sets of curves, one expects that the dis

crepancy must be due to circumstances not affecting the substance of the above theory.

The maxima of k and k/ν as function of ν do not greatly differ; thus it follows from our result that the maximum of absorption corresponds to the net creation of about 20 vibrational quanta for all temperatures. This, as we have mentioned, directly contradicts Muto's conclusion that the maximum of absorption corresponds to no change in the vibrational quanta; the nature of his argument is not clear to us, however, it appears that it must amount to ignoring the difference between $\bar{n} + 1$

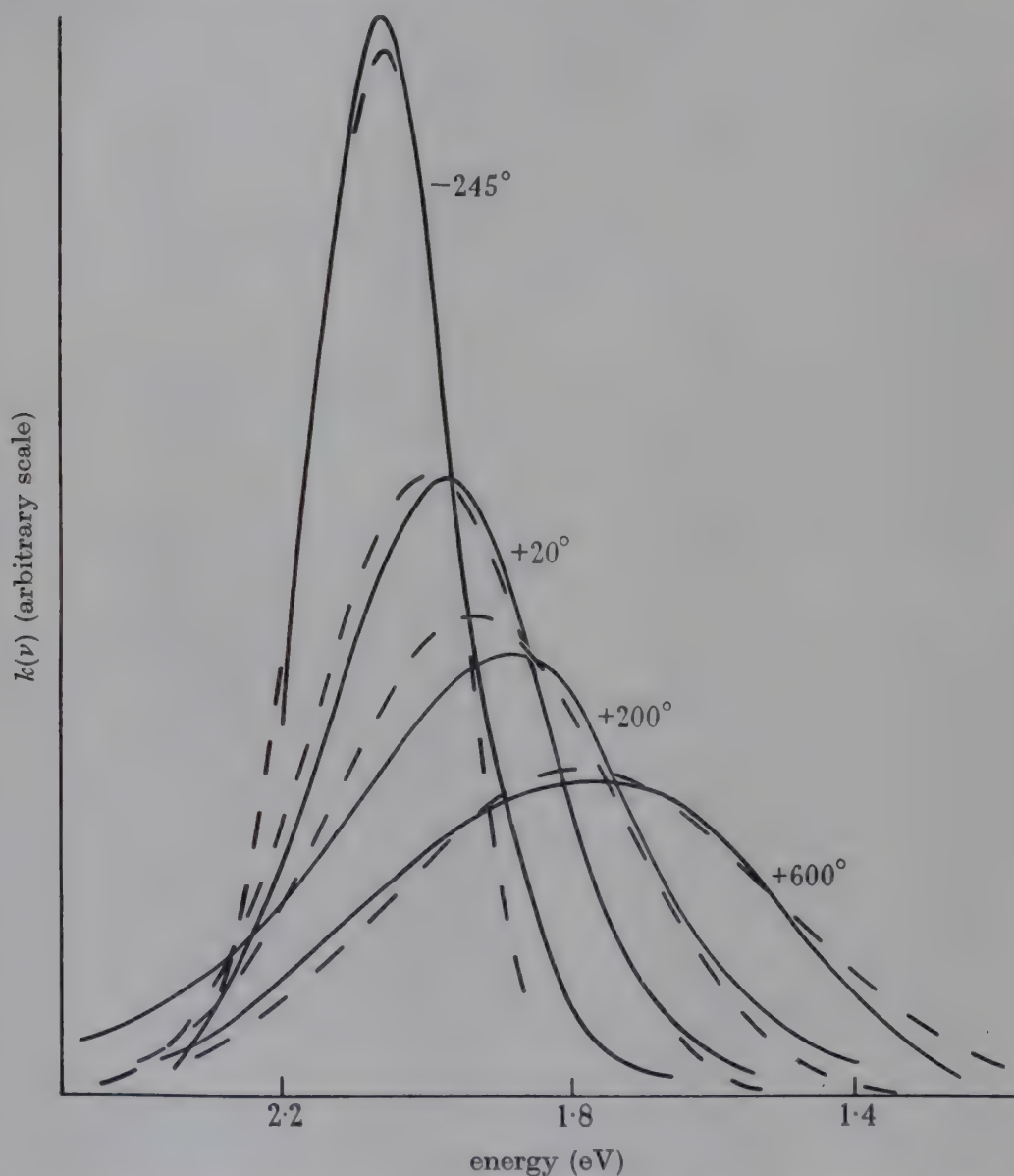


FIGURE 1. Theoretical and experimental absorption curves for F -centres in KBr. — experimental, ---- theoretical. $S = 22.4$; $\omega_l = 3.0 \times 10^{13}/\text{sec.}$; $E_{\mu^0\mu^0} = 1.67 \text{ eV}$; $B = -0.1 \text{ eV}$.

and $\bar{n}$, or the difference between the probabilities of an oscillator going up and down. Owing to his conclusion, however, he also finds it necessary to explain the shift of the absorption maximum with temperature; we believe that the explanation given by him correctly accounts for most of the discrepancy in the absolute positions in the absorption curves. We can put the idea as follows: when an F -centre changes its state, the lattice frequency is on the average altered by an amount of the order $1/N$,

although this does not appear in the harmonic forces approximation, which we have effectively used as regards the lattice. The total energy change associated with a transition is thus given by

$$\begin{aligned} h\nu &= E_{\mu''\mu'} + \Delta \sum_j (\bar{n}_j + \frac{1}{2}) \hbar\omega_l \\ &= E_{\mu''\mu'} + \hbar\omega_l \sum_j \Delta n_j + \sum_j (\bar{n}_j + \frac{1}{2}) B/N \\ &= E_{\mu''\mu'} + p\hbar\omega_l + B(\bar{n} + \frac{1}{2}). \end{aligned} \quad (5.2)$$

If this idea is correct, (5.2) should take the place of (4.5) as the relation between p and the absorption frequency ν .

With suitable choice of the constants $E_{\mu''\mu'}$ and B , we have plotted the theoretical curves for k on the absolute frequency scale in figure 1, in the same figure are also given the experimental curves of Pohl. The values used for the constants are given under the figure; ω_l has been calculated from the experimental values of $\epsilon_0, \epsilon_\infty, \omega_0$ (cf. Szigeti 1949). The agreement between the two sets of curves is seen to be satisfactory.

S is by far the most important constant in the theory. Once the wave functions of the F -centre electron in the ground and excited states are known, the electric fields $\mathcal{E}_{\mu'}, \mathcal{E}_{\mu''}$ can be calculated and hence also the value of S (cf. (3.7), (4.21)). We have made calculations of S , using wave functions given by Simpson (1949) for the NaCl lattice. Simpson considered two models for a positive impurity centre, namely, a vacant negative-ion site and a positive interstitial ion; the wave functions of the trapped electron are relatively more extended in the latter than the former. From the wave functions for the vacant negative-ion site, the value calculated for S is 3.6, which is small compared with the value 22.4 (for KBr!). Simpson, however, explained that the $2p$ wave function in this case is expected to be much too compact, therefore we repeated the calculation, using the $2p$ function for the interstitial ion model, with the result $S = 16$.

It is interesting to remark that theoretically an upper limit exists for the value of S . This corresponds to the case when the ground-state wave function is so localized that the field of the positive centre is completely screened off and the excited state wave function is so diffuse that the lattice is completely exposed to the field of the centre. The value of this upper limit is found to be $S = 55$.

6. PROBABILITY FOR NON-RADIATIVE TRANSITIONS

For the discussion of the non-radiative transitions, a closer description of the wave function for the complete system: F -centre + lattice is necessary. The natural characterization of the states does not correspond to true stationary states, but follows from the Born-Oppenheimer approximation, according to which the wave functions can be written in the following form:

$$\Psi_{\alpha,n}(\mathbf{x}, X, t) = \exp[-i\omega_{\alpha n}t] \psi_{\alpha}(\mathbf{x}, X) X_{\alpha n}(X), \quad (6.1)$$

where $\mathbf{x}, X$ stand respectively for the co-ordinates of the F -centre electron and the lattice particles, and α, n are the respective quantum numbers for the electronic motion in the F -centre and the lattice vibrations.

The Hamiltonian for the whole system can be written as

$$H(\mathbf{x}X) = K(\mathbf{x}) + V(\mathbf{x}) + K(X) + V(X) + V(\mathbf{x}X), \quad (6.2)$$

where $K(\mathbf{x})$ = kinetic energy of the F -centre electron,

$V(\mathbf{x})$ = interaction between the electron and the impurity centre,

$K(X)$ = kinetic energy of the lattice particles,

$V(X)$ = potential energy of the lattice,

$V(\mathbf{x}, X)$ = interaction between the F -centre and the lattice.

The Born-Oppenheimer wave functions satisfy the following equations:

$$\{K(\mathbf{x}) + V(\mathbf{x}) + V(\mathbf{x}X) - E_\alpha(X)\}\psi_\alpha(\mathbf{x}X) = 0, \quad (6.3)$$

$$\{K(X) + V(X) + E_\alpha(X) - E_{\alpha n}\}X_{\alpha n}(X) = 0. \quad (6.4)$$

(6.3) is the wave equation of the electron with the nuclei held in the configuration X . The corresponding eigenvalue $E_\alpha(X)$ is a function of X and acts as an effective potential for the lattice motion in equation (6.4).

As the states are not strictly stationary, transitions can occur between states of equal energy. As pointed out long ago by Kronig (1928), in cases such as this, when there is no perturbation term, the transition probabilities can still be obtained by perturbation methods. In fact, so long as the approximate wave functions are strictly orthogonal, one can apply Dirac's perturbation method in the usual way and show that the total transition probability from an initial state i to final states of approximately the same energy is given by

$$\frac{2\pi}{\hbar^2} \frac{1}{\Delta\nu} \sum_f^{0 < \nu_{fi} < \Delta\nu} |\langle f | H - E_i | i \rangle|^2, \quad (6.5)$$

where the summation over f is restricted by the frequency condition $0 < \nu_{fi} < \Delta\nu$ and E_i is the eigenvalue of state i .

In the present case, the matrix element in (6.5) can be expressed in a form suitable for calculation as follows: Multiplying (6.3) and (6.4) respectively by $X_{\alpha n}(X)$ and $\psi_\alpha(\mathbf{x}X)$ and adding the equations, one obtains

$$(H - E_{\alpha n})\psi_\alpha(\mathbf{x}X)X_{\alpha n}(X) = K(X)\psi_\alpha(\mathbf{x}X)X_{\alpha n}(X) - \psi_\alpha(\mathbf{x}X)K(X)X_{\alpha n}(X). \quad (6.6)$$

Writing $K(X)$ in terms of the normal co-ordinates q_j ,

$$K(X) = -\frac{1}{2}\hbar^2 \sum_j \frac{\partial^2}{\partial q_j^2},$$

and ignoring the second derivatives of $\psi_\alpha(\mathbf{x}X)$ with respect to q_j , one obtains from (6.6) the following expression for the required matrix element:

$$\langle \beta n' | H - E_{\alpha n} | \alpha n \rangle = -\hbar^2 \sum_j \int X_{\beta n'}^*(X) \frac{\partial}{\partial q_j} X_{\alpha n}(X) dX \left[\int \psi_\beta^*(\mathbf{x}X) \frac{\partial}{\partial q_j} \psi_\alpha(\mathbf{x}X) d\mathbf{x} \right]. \quad (6.7)$$

Following the same approximation as that already used in applying the Franck-Condon principle (Condon & Morse 1929), we shall take the electronic integral outside the integration over the lattice co-ordinates. If, for simplicity, we write

$$\frac{1}{\sqrt{N}} P_j(\beta\alpha) = \left\{ \int \psi_\beta^*(\mathbf{x}X) \frac{\partial}{\partial q_j} \psi_\alpha(\mathbf{x}X) d\mathbf{x} \right\}, \quad (6.8)$$

the transition probability (6.5) from the state α to the state β of the F -centre becomes

$$\frac{\hbar^2}{2\pi\Delta\nu} \sum_{n'}^{(0 \rightarrow \Delta\nu)} \left| \sum_j \frac{1}{\sqrt{N}} P_j(\beta\alpha) \int X_{\beta n'}^*(X) X_{\alpha n}(X) dX \right|^2, \quad (6.9)$$

where the summation over the final vibrational states is restricted by the transition frequency which must lie within the range $0 \rightarrow \Delta\nu$.

When the vibrational wave functions (3.12) are substituted in (6.9), we can once more consider all the transitions in which the number of oscillators going up exceeds the number going down by p and replace $\Delta\nu$ in (6.9) by $\omega_l/2\pi$; here as the transition frequency is in the neighbourhood of zero, so we are only concerned with the value of p given by

$$p\hbar\omega_l + E_{\beta\alpha} = 0. \quad (6.10)$$

The essential considerations required to work out (6.9) are the same as in §4; they are more involved, because the matrix element in (6.9) is a sum of overlap integrals rather than a single term as in §4. To save space, we shall not reproduce the arguments but write down the result directly:

$$\begin{aligned} \frac{\hbar^2}{\omega_l} \{ \frac{1}{2} Z^2 (\bar{n} R_{p+1} + (\bar{n} + 1) R_{p-1}) + |Y|^2 [(\bar{n} + \frac{1}{2})^2 + \frac{1}{2} \bar{n}(\bar{n} + 1)] R_p \\ - |Y|^2 (\bar{n} + \frac{1}{2}) [\bar{n} R_{p+1} + (\bar{n} + 1) R_{p-1}] + \frac{1}{4} |Y|^2 [\bar{n}^2 R_{p+2} + (\bar{n} + 1)^2 R_{p-2}] \}, \end{aligned} \quad (6.11)$$

where
$$R_p = \exp [- (2\bar{n} + 1) S] \left(\frac{\bar{n} + 1}{\bar{n}} \right)^{\frac{1}{2}p} I_p(2S \sqrt{[\bar{n}(\bar{n} + 1)]}), \quad (6.12)$$

$$Z^2 = \left(\frac{v_a \omega_l}{\hbar} \right) \sum_\lambda \int |P_{\mathbf{y}\lambda}(\beta\alpha)|^2 d\mathbf{y}, \quad (6.13)$$

$$Y = \frac{v_a}{\hbar\omega_l} \sum_\lambda \int P_{\mathbf{y}\lambda}(\beta\alpha) (A_{\mathbf{y}\lambda}^\beta - A_{\mathbf{y}\lambda}^\alpha) d\mathbf{y} \quad (6.14)$$

7. DISCUSSION OF THE NON-RADIATIVE TRANSITION PROBABILITIES

We shall ignore the small differences in the indices of R in (6.11). Thus we have

$$\frac{\hbar^2}{\omega_l} (\bar{n} + \frac{1}{2}) Z^2 R_p, \quad (7.1)$$

which we shall use as the basis for discussion in this section.

The value of Z^2 depends on the quantities $P_{\mathbf{y}\lambda}(\beta\alpha)$ defined in (6.8) which can be obtained readily with the help of the usual perturbation theory. To the normal modes are associated the electric field (Huang 1950):

$$\mathcal{E}(\mathbf{r}) = \omega_l \left[\frac{8\pi}{v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \right]^{\frac{1}{2}} \frac{1}{\sqrt{N}} \left\{ - \sum_{\mathbf{y}} q_{\mathbf{y}1} (\mathbf{y} / |\mathbf{y}|) \cos 2\pi \mathbf{y} \cdot \mathbf{r} + \sum_{\mathbf{y}} q_{\mathbf{y}2} (\mathbf{y} / |\mathbf{y}|) \sin 2\pi \mathbf{y} \cdot \mathbf{r} \right\}. \quad (7.2)$$

Taking this as a perturbing field on the F -centre electron and assuming it to be roughly uniform over the F -centre, one can immediately write down the first-order perturbed wave function for the F -centre electron in state α . It follows readily from the definition of $P_{\mathbf{y}\lambda}(\beta\alpha)$ that the coefficient of the β -component of the perturbed wave function of the α -state should be identical with

$$\frac{1}{\sqrt{N}} \sum_{\mathbf{y}\lambda} P_{\mathbf{y}\lambda} q_{\mathbf{y}\lambda}.$$

In this way one obtains

$$|P_{\mathbf{y}1}|^2 + |P_{\mathbf{y}2}|^2 = \omega_l^2 \left[\frac{8\pi}{v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \right] \frac{|\langle \beta | e\mathbf{x} \cdot \mathbf{y} | \alpha \rangle|^2}{|\mathbf{y}|^2 E_{\beta\alpha}^2}. \quad (7.3)$$

Z^2 defined in (6.13) is thus given approximately by

$$Z^2 = \frac{8\pi\omega_l^3}{3\hbar v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \frac{|\langle \beta | e\mathbf{x} | \alpha \rangle|^2}{E_{\beta\alpha}^2}. \quad (7.4)$$

Substituting (7.4) in (7.1), we obtain the following expression for the probability for an F -centre making a non-radiative transition from the state α to β :

$$\frac{16\pi^2\hbar\omega_l^2}{3v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \frac{|\langle \beta | e\mathbf{x} | \alpha \rangle|^2}{E_{\beta\alpha}^2} (\bar{n} + \frac{1}{2}) R_p. \quad (7.5)$$

It is our purpose to compare the non-radiative with the radiative transitions. The total radiative transition probability from an initial state i to final states f is given generally by

$$\frac{4}{3\hbar c^3} \sum_f |\langle i | e\mathbf{x} | f \rangle|^2 \omega_{fi}^3. \quad (7.6)$$

Following the same approximation as before, we can thus write the probability for an F -centre electron to return from the excited state μ'' to the ground state μ' as

$$\frac{4}{3\hbar c^3} |\langle \mu' | e\mathbf{x} | \mu'' \rangle|^2 \sum_{n''} \left| \int X_{\mu'n''}^*(X) X_{\mu''n''}(X) dX \right|^2 \omega_{\mu''n'',\mu'n'}^3. \quad (7.7)$$

When the vibrational wave functions (3.12) are substituted, the terms in the latter factor reduce to the overlap integrals already considered in §4. Since for all transitions in which the net number of vibrational quanta created is p , the radiation frequency $\omega_{\mu''n'',\mu'n'}$ has the same value:

$$\omega(p) = \omega_{\mu''n'',\mu'n'} = \frac{E_{\mu''\mu'}}{\hbar} - p\omega_l. \quad (7.8)$$

It follows immediately from the results obtained in §4 that the probability (7.7) can be written as

$$\frac{4}{3\hbar c^3} |\langle \mu' | e\mathbf{x} | \mu'' \rangle|^2 \sum_p R_p \omega^3(p). \quad (7.9)$$

R_p , we remember, has a maximum at $p \simeq 22$, (7.9) can thus be written approximately as

$$\frac{4}{3\hbar^4 c^3} |\langle \mu' | e\mathbf{x} | \mu'' \rangle|^2 [E_{\mu''\mu'} - 22\hbar\omega_l]^3 \sum_p R_p.$$

Using the series for I_p , it is easy to show that

$$\sum_{p=-\infty}^{\infty} R_p \equiv 1.$$

Hence the probability of radiative transitions is given approximately by

$$\frac{4}{3\hbar^4 c^3} | \langle \mu' | e\mathbf{x} | \mu'' \rangle |^2 [E_{\mu''\mu'} - 22\hbar\omega_l]^3. \quad (7.10)$$

An F -centre electron, raised to the excited state by light absorption, can either return to the ground state by (i) radiative transitions or (ii) non-radiative transitions, or (iii) be thrown into the conduction band by lattice vibrations. Mott (1938) has considered it possible that (ii) might jointly with (i) compete with the process (iii), which alone leads to photo-current. Thus let us consider first the relative probabilities of (i) and (ii). Putting in (7.5)

$$\beta = \mu', \quad \alpha = \mu'' \quad \text{and} \quad p = E_{\mu''\mu'}/(\hbar\omega_l) \quad (\text{cf. (6.10)})$$

and dividing the result by (7.10), we obtain the ratio of the probabilities for (ii) and (i):

$$\frac{(2\pi)^3 \hbar^5 \omega_l^2 c^3}{v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) \frac{(\bar{n} + \frac{1}{2}) R_p}{E_{\mu''\mu'}^2 (E_{\mu''\mu'} - 22\hbar\omega_l)^3}. \quad (7.11)$$

On putting in the numerical values (for KBr, $E_{\mu''\mu'}$ as given in figure 1), one finds

$$4 \times 10^5 (\bar{n} + \frac{1}{2}) R_{87}. \quad (7.12)$$

From the following values of this ratio:

T	-245°C	20°C	200°C	600°C
ratio (ii)/(i)	4×10^{-19}	2×10^{-9}	2×10^{-6}	0.8

it is seen that in the temperature range of interest, an excited F -centre returns to the ground state mainly by radiative transitions.

Consider next the relative magnitudes of (iii) and (ii). Let us represent a wave function of the electron in the conduction band approximately by

$$\exp(2\pi i \mathbf{k} \cdot \mathbf{r}). \quad (7.13)$$

The corresponding state has the energy $(\hbar^2 \mathbf{k}^2 / 2m)$ and will be specified by the index $\mathbf{k}$. Since an electron in state μ'' can be thrown into any state in the conduction band by absorbing the appropriate number of vibrational quanta, in order to obtain the probability for (iii), we have to put $\alpha = \mu''$, $\beta = \mathbf{k}$ in (7.5) and sum the expression over all states $\mathbf{k}$. Supposing that the excited state μ'' is W below the bottom of the conduction band, we have

$$E_{\mathbf{k}\mu''} = W + \frac{\hbar^2 \mathbf{k}^2}{2m}. \quad (7.14)$$

It follows thus that the probability for (iii) is

$$\frac{16\pi^2 \hbar \omega_l^2}{3v_a} \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_0} \right) (\bar{n} + \frac{1}{2}) \sum_{\mathbf{k}} \frac{| \langle \mathbf{k} | e\mathbf{x} | \mu'' \rangle |^2}{(W + \hbar^2 \mathbf{k}^2 / 2m)^2} R_p, \quad (7.15)$$

where p is related to k by (cf. (6.10))

$$p = -E_{\mathbf{k}\mu''}/(\hbar\omega_l). \quad (7.16)$$

If the $\mathbf{k}$ -state wave function used in evaluating the matrix element is normalized as in (7.13), the summation over $\mathbf{k}$ becomes simply an integration over $\mathbf{k}$ -space (i.e. with unit weight function), so that (7.15) becomes

$$\frac{16\pi^2\hbar\omega_l^2}{3v_a}\left(\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_0}\right)(\bar{n}+\frac{1}{2})\int\frac{|\langle\mathbf{k}|e\mathbf{x}|\mu''\rangle|^2}{(W+\hbar^2\mathbf{k}^2/2m)^2}R_p d\mathbf{k}. \quad (7.17)$$

The matrix element of the electric moment does not depend appreciably on $\mathbf{k}$, so long as the corresponding wave-length $1/|\mathbf{k}|$ is large compared with the μ'' -orbit; in fact, the other factor in the $\mathbf{k}$ -integral falls off with k so rapidly, only such values of $\mathbf{k}$ are relevant. Hence we write the probability as

$$\frac{16\pi^2\hbar\omega_l^2}{3v_a}\left(\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_0}\right)|\langle 0|e\mathbf{x}|\mu''\rangle|^2(\bar{n}+\frac{1}{2})\int\frac{R_p}{(W+\hbar^2\mathbf{k}^2/2m)^2}d\mathbf{k}, \quad (7.18)$$

$$\text{or} \quad \frac{8\pi}{3\hbar^4v_a}\left(\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_0}\right)|\langle 0|e\mathbf{x}|\mu''\rangle|^2(2m\hbar\omega_l)^{\frac{3}{2}}(\bar{n}+\frac{1}{2})\int_{-\infty}^{-p_0}\frac{(-p-p_0)^{\frac{3}{2}}}{p^2}dp, \quad (7.19)$$

where we have changed the integration variable from $\mathbf{k}$ to p with the help of (7.16) and p_0 is given by

$$p_0 = W/(\hbar\omega_l).$$

Dividing (7.19) by (7.10), we obtain the ratio of (iii) to (ii):

$$\frac{2\pi c^3}{v_a}\left(\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_0}\right)\frac{\left|\int\psi_{\mu''}\mathbf{x}d\mathbf{x}\right|^2}{\left|\int\psi_{\mu'}^*\mathbf{x}\psi_{\mu'}d\mathbf{x}\right|^2}\frac{(2m\hbar\omega_l)^{\frac{3}{2}}(\bar{n}+\frac{1}{2})}{[E_{\mu''\mu'}-22\hbar\omega_l]^3}\int_{-\infty}^{-p_0}R_p\frac{(-p-p_0)^{\frac{3}{2}}}{p^2}dp. \quad (7.20)$$

Assuming with Simpson (1949) the following radial functions respectively for the states μ' and μ'' :

$$\left(\frac{4\alpha_{\mu''}^5}{3}\right)re^{-\alpha_{\mu''}r}, \quad (4\alpha_{\mu'}^3)^{\frac{1}{2}}e^{-\alpha_{\mu'}r}, \quad (7.21)$$

one finds

$$\frac{2\pi c^3}{v_a}\left(\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_0}\right)\frac{(\alpha_{\mu'}+\alpha_{\mu''})^{10}}{4\alpha_{\mu'}^3\alpha_{\mu''}^{10}}\frac{(2m\hbar\omega_l)^{\frac{3}{2}}(\bar{n}+\frac{1}{2})}{[E_{\mu''\mu'}-22\hbar\omega_l]^3}\int_{-\infty}^{-p_0}R_p\frac{(-p-p_0)^{\frac{3}{2}}}{p^2}dp. \quad (7.22)$$

When the values (Simpson 1949)

$$\alpha_{\mu''} \simeq \frac{1}{2}\alpha_{\mu'} = 0.66 \text{ atomic units,}$$

for $\alpha_{\mu''}, \alpha_{\mu'}$ as well as the numerical values for the other constants are put in, (7.22) becomes

$$1.7 \times 10^9 (\bar{n} + \frac{1}{2}) \int_{-\infty}^{-p_0} R_p \frac{(-p-p_0)^{\frac{3}{2}}}{p^2} dp. \quad (7.23)$$

In order to estimate this ratio, the coupling constant S for the transition from μ'' to the ionized state has to be known. We have tentatively used the value for the $\mu' \rightarrow \mu''$ transition, namely, 22.4. The logarithm of the ratio calculated is plotted against $1/T$ for $W = 0.1, 0.2$ and 0.4 eV in figure 2. Apart from the highest temperatures, the curves are practically straight. The curve for $W = 0.1$ eV cuts the axis at $\sim 130^\circ \text{ K}$ (or $\sim -140^\circ \text{ C}$). At this temperature, the probabilities of (i) and (ii) are

comparable; but at $\sim 100^\circ\text{K}$ the non-radiative transitions have fallen by a factor ~ 1000 . Thus the situation represented by this curve resembles quite closely the observed course of photo-current in NaCl and KCl (Pohl 1937). It is however to be borne in mind that the value for S used in calculating (iii) is only very tentative.

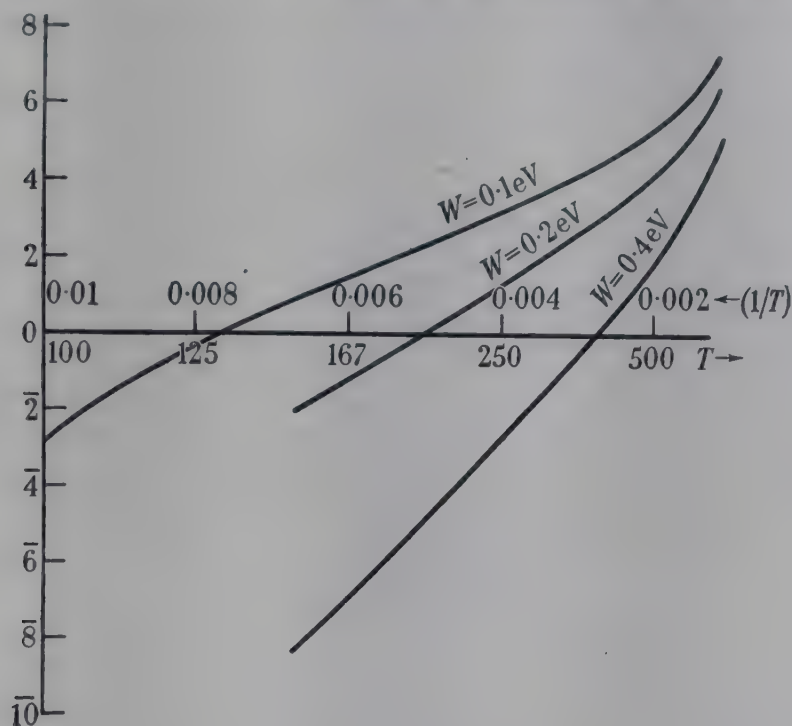


FIGURE 2. Logarithm of the ratio of the non-radiative to radiative transitions as function of $1/T$ (excited level of F -centre W below bottom of conduction band).

REFERENCES

- Condon, E. U. & Morse, P. M. 1929 *Quantum mechanics*. New York: McGraw-Hill.
 Finkelburg, W. 1938 *Kontinuierliche Spektren*. Berlin: Springer.
 Frenkel, J. 1936 *Phys. Z. Sowjet.* **9**, 158.
 Gurney, R. W. & Mott, N. F. 1938 *Trans. Faraday Soc.* **34**, 506.
 von Hippel, A. 1936 *Z. Phys.* **101**, 680.
 Huang, K. 1950 Report L/T239 of British Electrical and Allied Industries Research Association.
 Kronig, R. de L. 1928 *Z. Phys.* **50**, 347.
 Mott, N. F. 1938 *Proc. Roy. Soc. A*, **167**, 384.
 Mott, N. F. & Gurney, R. W. 1948 *Electronic processes in ionic crystals*, 2nd ed. Oxford: Clarendon Press.
 Muto, T. 1949 *Progr. Theor. Phys. (Japan)*, **4**, 181.
 Peierls, R. 1932 *Ann. Phys., Lpz.*, **13**, 905.
 Pohl, R. W. 1937 *Proc. Phys. Soc.* **49** (extra part), 3.
 Seitz, F. 1939 *Trans. Faraday Soc.* **35**, 74.
 Simpson, J. H. 1949 *Proc. Roy. Soc. A*, **197**, 269.
 Szigeti, B. 1949 *Trans. Faraday Soc.* **45**, 155.
 Whittaker, E. T. & Watson, G. N. 1946 *Modern analysis*, 4th ed. Cambridge University Press.

Ultrasonic dispersion in organic vapours

BY J. D. LAMBERT AND J. S. ROWLINSON

Physical Chemistry Laboratory, University of Oxford

(Communicated by Sir Cyril Hinshelwood, F.R.S.—Received 8 August 1950)

The velocity of ultrasonic waves has been measured in a series of organic vapours at frequencies 566, 1192 and 4000 kc./sec., temperatures 20 and 100° C, and pressures ranging from 0.25 to 4 atm. Ultrasonic dispersion was found with benzene, cyclopropane and ethane. No dispersion was found with propane, *n*-hexane, cyclohexane, cyclohexene, ethyl chloride, chloroform, acetonitrile, acetaldehyde, acetone, diethyl ether and methyl alcohol. The mechanism and rate of the activation of vibrational energy by intermolecular collisions are discussed in relation to molecular structure. It is concluded that, for polyatomic molecules in general, there is little hindrance to the interconversion of translational and vibrational energy, especially at high temperatures. Slow interconversion, leading to ultrasonic dispersion, only occurs with small or rigid molecules, where internal rotations or vibrations of low frequency are absent. Interchange of vibrational energy between the different modes within the molecule is usually very rapid.

INTRODUCTION

Measurements of ultrasonic dispersion in gases have established that, for a number of simple molecules such as chlorine, carbon dioxide, nitrous oxide, methane, the transfer of energy between translational and vibrational degrees of freedom is an inefficient process. Several thousand intermolecular collisions are necessary for each quantum of vibrational energy acquired or lost, so that a vibrationally activated molecule may have a considerable 'collision life'. Such persistence of vibrational energy is an important factor in the kinetics of gas reactions and in combustion theory in particular. Systematic measurements have therefore been made on a number of simple hydrocarbons and organic vapours, in order to discover whether they behave similarly to the diatomic and triatomic molecules previously investigated, and whether any connexion can be established between structural features and the ease of acquiring or losing vibrational energy.

EXPERIMENTAL

Apparatus

Two acoustic interferometers were used:

(1) The instrument previously described (Alexander & Lambert 1942), using piezoelectric quartz oscillators of frequency 566 and 1192 kc./sec. driven by the circuit described by Eucken & Jaacks (1935).

(2) A high-frequency interferometer using a 4 Mc./sec. quartz crystal. This was based on the design of Stewart (1946) and is shown in figure 1. The quartz is an optically flat disk approximately 0.5 mm. thick and 10 mm. in diameter, gold plated on both faces. The reflector is a bronze piston moving in a thick brass cylinder. The piston head is polished flat, and is accurately parallel with the under surface of the quartz: a small groove is cut in its side to allow gas to flow freely past. The piston is

moved by an Invar rod 20 cm. long, whose lower end is held in contact with a micrometer head by a helical spring. The thrust of the spring can be adjusted to compensate for the effect of altering the pressure of the gas in the interferometer. The micrometer head is graduated to 0.0002 cm. and can be read to half a division. The

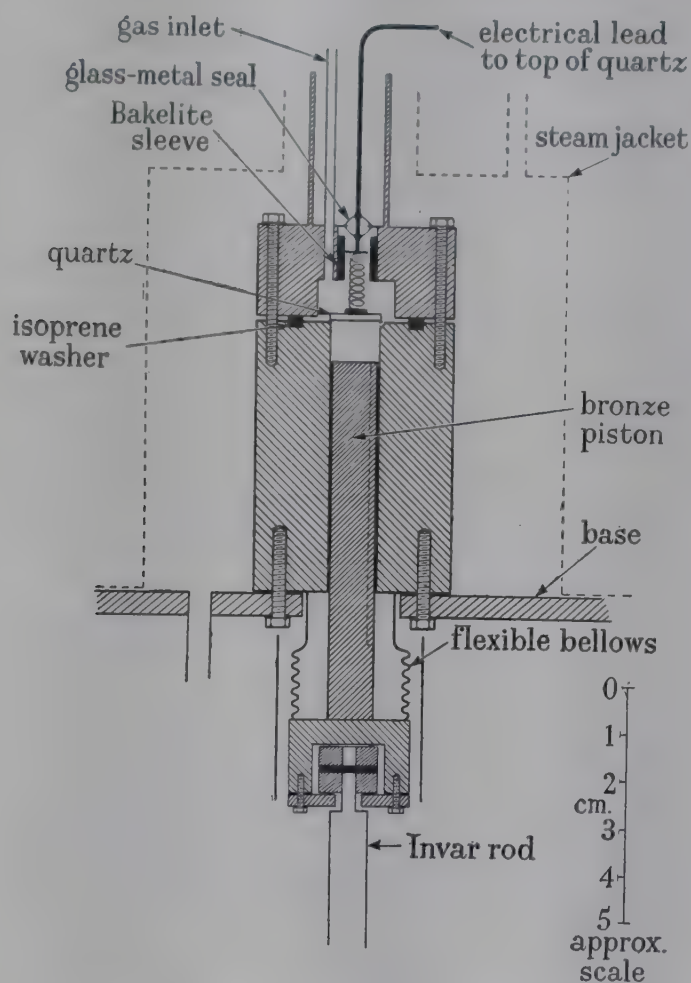


FIGURE 1. High-frequency interferometer.

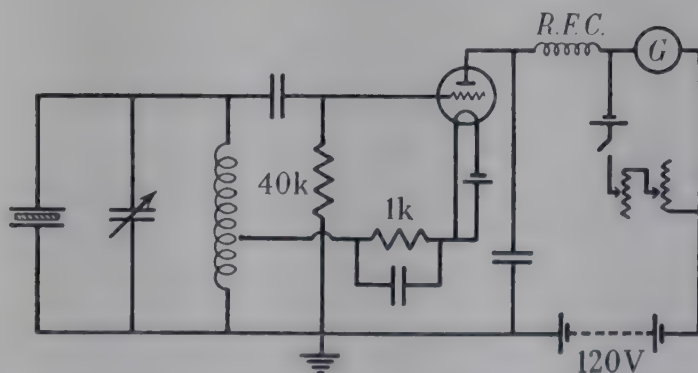


FIGURE 2. Diagram of 4 Mc./sec. oscillator circuit.

quartz is driven by a variable frequency oscillator circuit (figure 2) due to Carlin (1949). The fluctuations in anode current produced by varying the distance between reflector and quartz are measured by a galvanometer balanced in a potentiometer circuit. The positions corresponding to standing waves are marked by sharp minima in the galvanometer deflexion, and can be read on the micrometer head. The

measured wave-lengths ranged between 0.005 and 0.007 cm.; results obtained by averaging over about twenty minima were reproducible to 1 part in 300. The wave-length is so small in comparison with the apparatus dimensions that no tube correction is necessary.

The whole interferometer is vacuum-tight, and is connected through a metal-glass joint to a vacuum and vapour supply system, as described previously. When measurements are made on condensible vapours, the interferometer is heated by a steam jacket and all connecting tubing heated electrically by a Nichrome winding. This apparatus is used at pressures up to 1 atm. only, and glass taps lubricated with Silicone grease are used. Pressure is measured by a steam jacketed mercury U-tube manometer connected to the interferometer through an efficient gold trap, to prevent diffusion of mercury vapour. The pressure in the interferometer system is not measurably altered by the small movements of the piston involved in making measurements.

Materials

Most of the materials were purified and introduced into the apparatus by methods described previously (Lambert, Roberts, Rowlinson & Wilkinson 1949; Lambert, Staines & Woods 1950). In addition:

Propane, obtained from a cylinder, was passed through Sofnolite and calcium chloride tubes, and distilled *in vacuo* four times.

Cyclopropane was obtained from a cylinder supplied for anaesthetic purposes. The batch analysis was stated to show 99.4 % purity.

Cyclohexene was freed from peroxide by shaking with saturated ferrous sulphate solution. It was then washed with water, dried over calcium chloride, fractionated, and stored over sodium wire in an atmosphere of nitrogen.

Procedure

Wave-lengths were obtained by comparison with measurements on pure, dry air, free from carbon dioxide, which were made under identical conditions with the measurements on the vapour under investigation. Measurements were made at several pressures for each vapour. For any imperfect gas both the velocity of sound W and the apparent heat capacity C_r vary with pressure in a manner depending on the second virial coefficient B , and must be corrected to zero pressure to give W_0 and C_{r_0} . In addition, for a dispersing gas, both W_0 and C_{r_0} (corrected for gas imperfection only) vary with ν/p (cf. Eucken & Becker 1934), where ν is the sound frequency and p the pressure at which W is measured. For the dispersing vapours, gas imperfection corrections were applied and values of C_{r_0} calculated for each pressure by the method described previously (Alexander & Lambert 1942). For non-dispersing vapours, where C_{r_0} is constant and independent of ν/p , the method of Itterbeek & Doninck (1946) was used.

The variation of sound velocity with pressure is here given by

$$W = W_0 \left(1 + \frac{S}{RT} p \right), \quad (1)$$

where S is a function of the second virial coefficient B given by

$$S = B + \frac{T}{c} \left(\frac{dB}{dT} \right) + \frac{T^2}{2c(c+1)} \left(\frac{d^2B}{dT^2} \right), \quad (2)$$

where $c = C_{v_0}/R$. A plot of W against p therefore gives a straight line, which may be extrapolated to zero pressure to give W_0 . This method has the advantage that accurate knowledge of the second virial coefficient and its derivatives is not required. There is no risk of applying it inadvertently in a dispersion region, as the slope of the line then changes rapidly, and, even if the plot is apparently linear, impossible values are obtained for C_{v_0} .

RESULTS

The results obtained with non-dispersing vapours are summarized in table 1. Observed values of C_{v_0} are compared with the best available values from other sources, either experimental or spectroscopic. Agreement is on the whole good. Since the velocity of sound varies with $\sqrt{\gamma}$, it is not to be expected that observed values of C_{v_0} which lie as high as 30 cal./mol. should be very accurate. There is no doubt that dispersion is absent in these cases, since this would result in a massive and rapid decrease in C_{v_0} towards 6 cal./mol. with rise in ν/p . For some cases the measured values of the function S are compared with values calculated from second virial coefficients measured in this laboratory (Lambert *et al.* 1949). Agreement is within the limits of experimental error (about ± 50 ml./mol.) in most cases. The general results for acetaldehyde differ from those reported previously (Alexander & Lambert 1942) for reasons which are discussed in the following section.

The observed values of C_{v_0} for benzene vapour at 117° C are shown plotted against $\log \nu/p$ in figure 3. Correction for gas imperfection was made by the Berthelot equation (cf. Lambert *et al.* 1949). Very strong dispersion accompanied by absorption

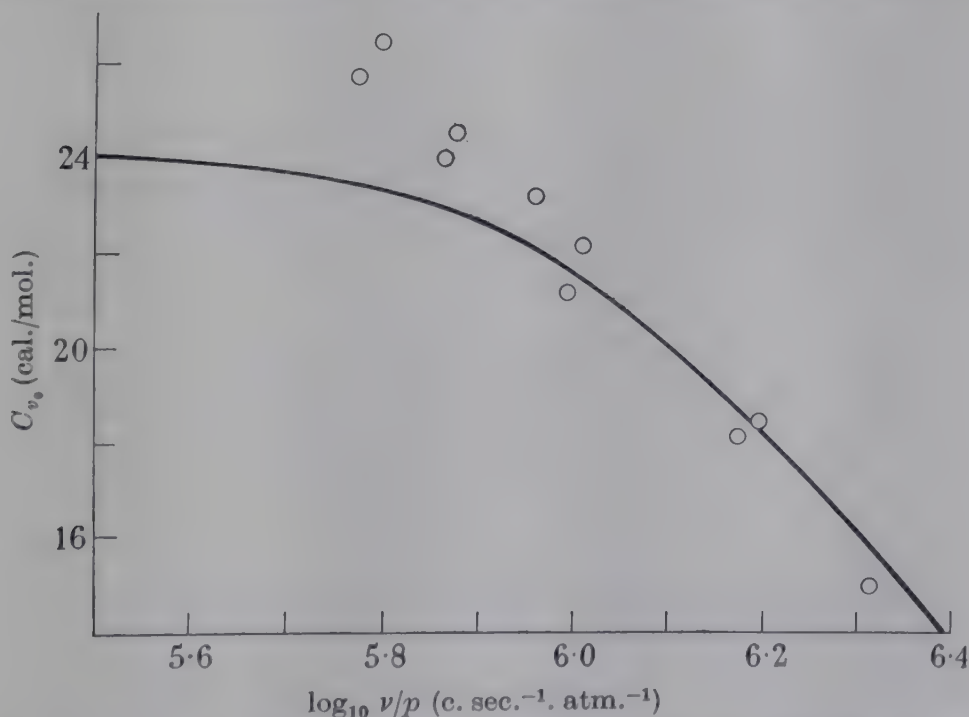


FIGURE 3. Benzene. O observed values at 117° C. — theoretical curve for $\beta = 1.33 \times 10^{-7}$ sec. Normal value of C_{v_0} at 117° C = 24.2 cal./mol. (Pitzer & Scott 1943).

TABLE I

substance	ν (kc./sec.)	pressure range (atm.)	temp. (° C)	C_v obs. (cal./mol.)	C_v other sources (cal./mol.)	$-S_{\text{obs.}}$ (ml./mol.)	$-S_{\text{calc.}}$ (ml./mol.)	dispersion
ethane	1192	0.2 to 4.9	20.3	10.8	10.5 (a)	100	140	—
	4000	0.15 to 1.0	19.6	10.8	10.5	190	140	+ below 0.3 atm.
propane	4000	0.25 to 1.0	19.6	15.8	15.3 (b)	300	335	—
<i>n</i> -hexane	1192	0.5 to 1.0	108.6	35	39.5 (c)	800	840	—
	4000	0.3 to 0.8	100.0	38	39.0	—	—	—
cyclohexane	1192	0.2 to 0.8	109.0	32	32.2 (d)	780	790	—
	4000	0.2 to 0.8	100.0	28	31.4	—	—	—
cyclohexene	1192	0.3 to 1.0	105.0	28	30.6 (d)	780	840	—
	4000	0.5 to 0.8	100.0	28	30.2	—	—	—
ethyl chloride	4000	0.2 to 0.5	19.6	12.8	12.4 (e)	—	—	—
	566	0.3 to 1.2	109.8	15.0	15.6 (f)	580	470	—
chloroform	1192	0.7 to 1.3	110.5	15.6	15.6	—	—	—
acetonitrile	566	0.2 to 0.6	105.8	11.4	11.9 (g)	1300	1220	—
	4000	0.2 to 0.8	100.0	11.0	11.8	1210	1220	—
acetaldehyde	566	0.2 to 0.9	20.1	11.1	10.8 (h)	830	855	—
	566	0.2 to 0.8	110.0	13.8	13.3	350	350	—
	1192	0.2 to 1.0	22.0	10.2	10.8	770	855	—
	1192	0.3 to 1.7	113.0	13.2	13.4	380	350	—
	4000	0.2 to 0.9	20.6	11.4	10.8	830	855	—
acetone	566	0.2 to 2.3	112.6	17.5	19.8 (j)	680	640	—
	1192	0.2 to 1.5	110.1	17.5	19.8	680	640	—
diethyl ether	1192	0.3 to 2.7	109.5	30.0	30.6 (k)	500	510	—
methyl alcohol	566	0.3 to 0.9	133.4	10.3	10.7 (l)	320	270	—
	1192	0.3 to 1.3	110.0	10.0	10.2	380	360	—
	4000	0.3 to 1.0	100.0	10.2	10.0	470	390	—

(a) Kistiakowsky & Rice (1939).

(b) Kistiakowsky & Rice (1940).

(c) Waddington & Douslin (1947).

(d) Montgomery & de Vries (1942).

(e) Eucken & Franck (1948).

(f) Vold (1935), *spectroscopic*.(g) Herzberg (1945), *spectroscopic*.

(h) Pitzer & Weltner (1949).

(j) Collins, Coleman & de Vries (1949) (corrected for gas imperfection by Lambert *et al.* 1949).

(k) Jennings & Bixler (1934) (corrected for gas imperfection as (j)).

(l) See Rowlinson (1948).

sets in at a value of $\log \nu/p = 5.7$, and C_v has fallen from 25 to 15 cal./mol. when $\log \nu/p = 6.3$. Absorption at 1192 and 4000 kc./sec. was so strong that the vapour was apparently opaque to sound, and no measurements could be made at any pressure. The dispersion region is therefore incomplete when $\log \nu/p = 7.2$. Measurements made at 96.5 and 165° C showed similar behaviour.

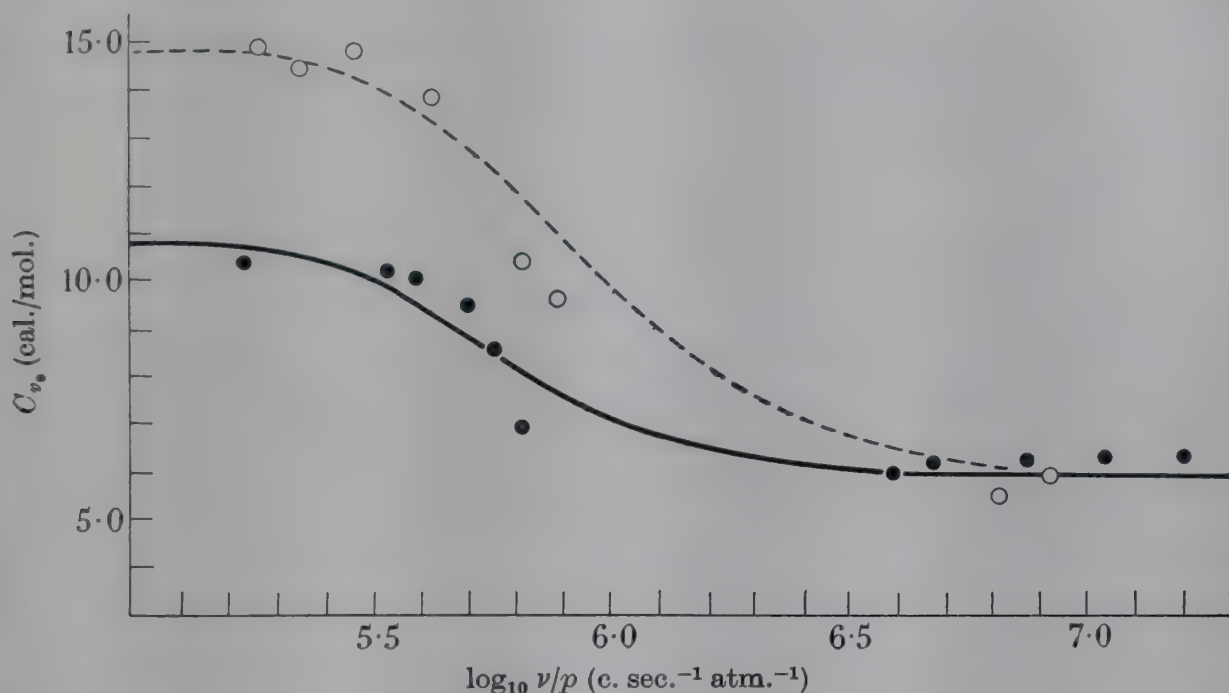


FIGURE 4. Cyclopropane. $\circ$ observed values at 100° C. $\bullet$ observed values at 18° C. theoretical curve for $\beta = 2.70 \times 10^{-7}$ sec. — theoretical curve for $\beta = 3.64 \times 10^{-7}$ sec. Normal value of C_v at 100° C = 15.0 cal./mol., at 18° C = 11.0 cal./mol. (Kistiakowsky & Rice 1940).

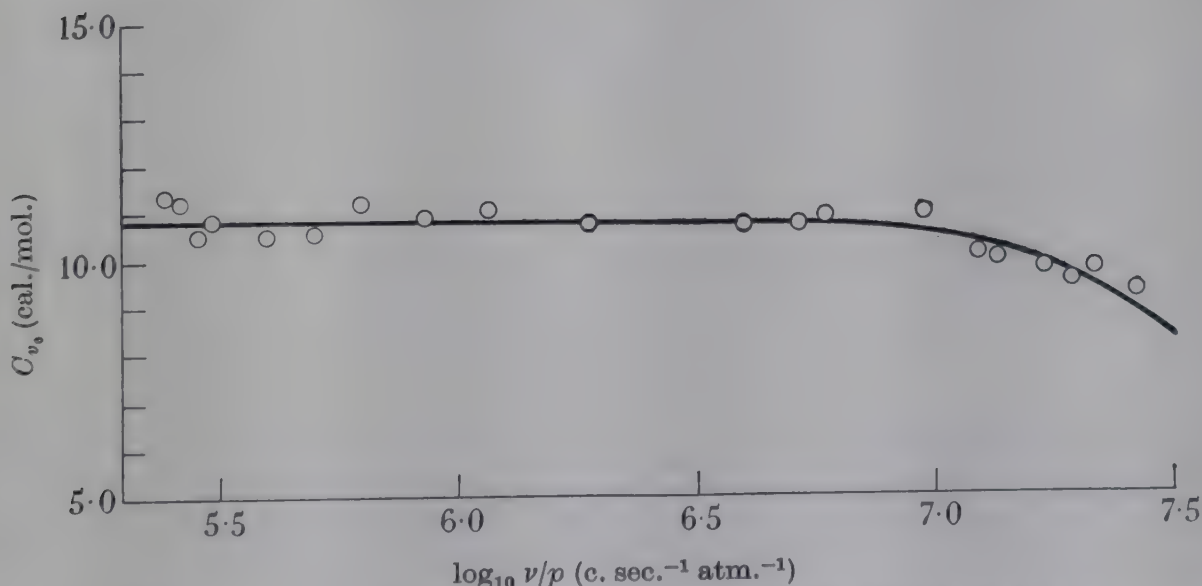


FIGURE 5. Ethane. $\circ$ observed values at 20° C. — theoretical curve for $\beta = 5.24 \times 10^{-9}$ sec. Normal value of C_v at 20° C = 10.5 cal./mol. (Kistiakowsky & Rice 1939).

The results for cyclopropane at 18 and 100° C are plotted in figure 4. Correction for gas imperfection was made by the Berthelot equation, taking $T_c = 107^\circ$ C and $p_c = 45$ atm. by analogy with propane. Cyclopropane would be expected to obey the

Berthelot equation at the temperatures used (Lambert *et al.* 1949), and a single compressibility measurement made at 78°C showed agreement with the calculated value within the limits of experimental error. Dispersion sets in at a value of $\log \nu/p = 5.2$ and is complete when $\log \nu/p = 7.0$, where the value of C_{v_0} at both temperatures is 6.0 cal./mol., corresponding to a gas with translational and rotational energy only.

The results for ethane are plotted in figure 5. Correction for gas imperfection was made by the Berthelot equation (Lambert *et al.* 1949). Dispersion appears to set in at the extreme end of the range investigated, where $\log \nu/p = 7.0$.

DISCUSSION

Dimerization dispersion

There is evidence that the more strongly polar vapours, acetaldehyde, acetone, acetonitrile and methyl alcohol, are slightly associated (Lambert *et al.* 1949, 1950). This gives rise to an increase in the compressibility of the vapour, which appears in the observed value of the second virial coefficient B , and hence in the observed value of S . (C_{v_0} , which is extrapolated to zero pressure, will be unaffected.) The possibility arises that the rate of association or dissociation may be so slow that there is not time for equilibrium to be established during the period between adiabatic compression and expansion when a sound wave travels through the vapour. The observed value of S would then be expected to show a frequency dispersion.

The theory of sound velocity in a dimerizing gas has been treated by Einstein (1920), on the following basic assumptions:

For a reversible dimerization process $2A \rightleftharpoons A_2$, let there be n_1 mol. of A and n_2 of A_2 per ml., and let the rate of association be

$$k_1 \left(\frac{n_1}{v} \right)^2$$

and the rate of dissociation

$$k_2 \left(\frac{n_2}{v} \right).$$

Then the mass-action equilibrium constant

$$K_c = \frac{n_1^2}{n_2 v} = \text{const. } e^{-\Delta U/RT},$$

where ΔU is the heat of association at constant volume. It is assumed that the molar heat capacity of the dimer is twice that of the monomer, and an expression for the velocity of sound in a perfect gas which undergoes dimerization is derived in terms of the sound frequency and the pressure. Einstein's treatment may be adapted to show how, for a slightly dimerized perfect gas, the function S will change with frequency. It has been shown (Lambert *et al.* 1949) that dimerization of this kind gives rise to an apparent second virial coefficient

$$B_d = \frac{-RT}{K_p} = \frac{-1}{K_c}.$$

The Einstein equations may then be rearranged to give

$$S_d = \frac{B_d}{2} \left[1 + \left(1 + \frac{2\Delta U}{C_v T} + \frac{\Delta U^2}{C_v C_p T^2} \right) (1 + \tau^2 \omega^2)^{-1} + \frac{R \tau^2 \omega^2}{C_p} (1 + \tau^2 \omega^2)^{-1} \right], \quad (3)$$

where $\tau = 1/k_2$, $\omega = 2\pi\nu$, and C_p is the molar heat capacity of the monomer. When $\tau\omega \rightarrow 0$, this reduces to

$$S_d^0 = B_d + \frac{T}{c} \left(\frac{dB_d}{dT} \right) + \frac{T^2}{2c(c+1)} \left(\frac{d^2 B_d}{dT^2} \right) \quad (4)$$

(compare equation (2) above). When $\tau\omega \rightarrow \infty$

$$S_d^\infty = \frac{B_d}{2} \left(1 + \frac{R}{C_p} \right). \quad (5)$$

S_d and B_d represent the effect of dimerization only, in an otherwise perfect gas (i.e. a gas for which, as far as gas imperfection is concerned, B and S are both zero). For a real dimerizing vapour the effects of gas imperfection and dimerization may be regarded as superposed (Lambert *et al.* 1949; Rowlinson 1949), so that

$$B_{\text{obs.}} = B_d + B, \quad S_{\text{obs.}} = S_d + S.$$

It may be assumed that B and S are frequency independent, at the sound frequencies used, so that the variation in $S_{\text{obs.}}$ with sound frequency will be given by equations (3), (4) and (5).

Acetaldehyde and acetonitrile have been shown to undergo simple dimerization, and measured values of K_p and B_d are available for insertion into equations (4) and (5). In both cases the calculated difference between S_d^0 and S_d^∞ is found to be so small as to lie entirely within the range of likely experimental error. There can therefore be no measurable dispersion in sound *velocity* due to the dimerization process. This is confirmed by the observed values of S recorded in table 1; there is no definite trend of the values obtained with different frequencies, and there is fair agreement with values calculated from measured second virial coefficients. There seems to be some evidence that upsetting of the dimerization equilibrium by the passage of sound waves may give rise to abnormal sound *absorption* in a limited frequency range, and this is being further investigated.

Earlier results on acetaldehyde (Alexander & Lambert 1942) were calculated on the assumption that the dimerization equilibrium does not follow the adiabatic compression cycle, and that the theoretical value of B for the *non-associated* vapour (Berthelot equation) should be used for the gas imperfection correction. This is incorrect, and gave rise to apparent multiple dispersion regions for C_{v_0} , with steps corresponding to the different frequencies used. Recalculation of these results, confirmed by further measurements, has shown that dispersion does not occur at all. The observed values of S correspond with the *measured* values of the second virial coefficient, and the values of C_{v_0} obtained are in agreement with the accurate thermal values now available (Pitzer & Weltner 1949).

The collision life of vibrational energy

The calculation of the time of relaxation β for interconversion of translational and vibrational energy in molecules where more than one vibration is active has been

treated theoretically by Schäfer (1940). In cases where only one vibration is active, giving a single time of relaxation, the variation of C_{v_0} with circular frequency is given by

$$C_{v_0}^\omega = \frac{U^2 + V^2}{U}, \quad (6)$$

$$U = C_0 + \frac{C_s}{1 + \beta^2 \omega^2}, \quad V = \frac{C_s \beta \omega}{1 + \beta^2 \omega^2}, \quad (7)$$

where C_s is the vibrational heat capacity, and C_0 the translational + rotational heat capacity.

This reduces to

$$C_{v_0}^\omega = \frac{1}{1 + \beta \omega} \left[C_{v_0} + C_0 \beta^2 \omega^2 + \frac{C_s^2 \beta^2 \omega^2}{C_{v_0} + C_0 \beta^2 \omega^2} \right]. \quad (8)$$

Since changes in $\nu (= \omega/2\pi)$ are exactly equivalent to changes in $1/p$ (if collision activation is regarded as a bimolecular process), equation (8) represents the variation of C_{v_0} with ν/p .

Where more than one vibrational mode is active, two limiting cases are considered:

I. The different modes are independently activated in collisions, and have different times of relaxation $\beta_1, \beta_2, \beta_3$, etc. This gives rise to a different expression for U and V in equation (6), leading to a broader dispersion zone. Large differences in β_1, β_2 would lead to disconnected dispersion zones.

II. One mode is activated directly with a relaxation time β_1 . The energy spreads from this to other modes with relaxation times β_{12}, β_{13} , etc. This again gives rise to a different expression for U and V . If, however, β_{12}, β_{13} , etc., are negligibly small in comparison with β_1 (i.e. the energy flows very easily from the first mode into the other modes), the equations reduce to equation (8), with a value of $\beta = \beta_1(C_s/C_1)$, where C_1 is the heat capacity of the vibrational mode which is activated directly by collision.

Examination of the experimental results for cyclopropane, which are plotted in figure 4, shows that the width of the dispersion zone corresponds to the theoretical curve calculated from equation (8) for a *single* β value. This points to the inference that vibrational activation is by mechanism II with $\beta_{12} \ll \beta_1$. With benzene and ethane the results do not cover the whole dispersion zone, but, as shown in figures 3 and 5, the points correspond reasonably well with theoretical curves calculated for a single β value.

Values of β calculated in this way for a pressure of 1 atm. for the hydrocarbons investigated are shown in table 2. Results for methane (Eucken & Aybar 1940), ethylene (Richards & Reid 1934), and propylene (Telfair 1942) are also included. β is related to the number of collisions necessary for a molecule to lose one quantum of vibrational energy Z_{10} by the equation

$$Z_{10} = Z\beta(e^{-h\nu/kT} + 1),$$

where Z is the number of collisions one molecule suffers per second (cf. Eucken & Becker 1934). If $h\nu \gg kT$, $Z^* = Z\beta$ will be approximately equal to Z_{10} . Values of Z^* are given in table 2, which have been calculated from the usual formula for Z , taking the collision diameters of Hirschfelder, Bird & Spotz (1949). The last

column of table 2 contains values of $\nu_{\min.}$, the probable lowest vibration frequency of the molecule.

TABLE 2

substance	temp. (° C)	β (sec.)	Z^*	$\nu_{\min.}$ (cm. ⁻¹) (a)
methane	109	8.4×10^{-7}	7600	1306
	200	6.7×10^{-7}	5400	—
	322	5.1×10^{-7}	3300	—
	353	4.9×10^{-7}	2950	—
ethylene	30	2.38×10^{-7}	2240	825
cyclopropane	18	3.64×10^{-7}	4000	740
	100	2.70×10^{-7}	2570	—
benzene	96	1.50×10^{-7}	1070	405
	117	1.33×10^{-7}	950	—
	165	1.18×10^{-7}	840	—
ethane	20	5.24×10^{-9}	52	275
propane	20	$< 10^{-8}$	< 90	?202
propylene	0-217	$< 10^{-8}$	< 90	?177
n-hexane	100	$< 10^{-8}$	< 90	—
cyclohexane	100	$< 10^{-8}$	< 90	—
cyclohexene	100	$< 10^{-8}$	< 90	—

(a) Herzberg (1945).

Inspection of table 2 shows that the values of Z^* run roughly parallel to the values of $\nu_{\min.}$. This is in accord with the view that vibrational activation takes place via the mode of lowest frequency, which will be the mode most easily activated. All the molecules which show dispersion are comparatively rigid, whilst those which do not, possess either internal rotations or ring deformation vibrations of very low frequency. All the substances listed in table 1, which do not show dispersion, have molecules with either internal rotations or low minimum vibration frequencies. It is interesting to compare methane ($\nu_{\min.} = 1306$) which shows dispersion, with chloroform ($\nu_{\min.} = 260$) which does not. If the mechanism of activation which has been suggested is correct, the values of Z^* are, in fact, too large by factors of possibly 2 or 3, since Z^* should be calculated from the time of relaxation for the mode first activated, β_1 , and this is equal to $\beta \times C_1/C_s$ (see above). Since $\beta_{12} \ll \beta_1$, vibrational energy must flow very rapidly from the mode first activated into the other modes in most of the molecules investigated. In all cases where measurements have been made at more than one temperature, Z^* and β decrease with rise in temperature.

The conclusion may be drawn that, for polyatomic molecules in general, there is little hindrance to the interconversion of translational and vibrational energy, especially at high temperatures. The 'persistence' of a quantum of vibrational energy (Z^*) is usually considerably less than 90 intermolecular collisions. Only in small or rigid molecules, where internal rotations or vibrations of low frequency are absent, is there high persistence, leading to values of Z^* rising above 1000 collisions. Interchange of vibrational energy between the different modes within the molecule is usually very rapid.

One of the authors, J.S.R., is indebted to the Department of Scientific and Industrial Research for a maintenance allowance.

REFERENCES

- Alexander, E. A. & Lambert, J. D. 1942 *Proc. Roy. Soc. A*, **179**, 499.
 Carlin, B. 1949 *Ultrasonics*. New York: McGraw-Hill.
 Collins, B. T., Coleman, C. F. & de Vries, T. 1949 *J. Amer. Chem. Soc.* **71**, 2929.
 Einstein, A. 1920 *S.B. preuss. Akad. Wiss.*, p. 380.
 Eucken, A. & Aybar, S. 1940 *Z. phys. Chem. B*, **46**, 195.
 Eucken, A. & Becker, R. 1934 *Z. phys. Chem. B*, **27**, 235.
 Eucken, A. & Franck, E. U. 1948 *Z. Elektrochem.* **52**, 195.
 Eucken, A. & Jaacks, H. 1935 *Z. phys. Chem. B*, **30**, 85.
 Herzberg, G. 1945 *Infrared and Raman spectra of polyatomic molecules*. New York: van Nostrand.
 Hirschfelder, J. O., Bird, R. B. & Spotz, E. L. 1949 *Chem. Rev.* **44**, 205.
 van Isterbeek, A. & van Doninck, W. 1946 *Proc. Phys. Soc.* **58**, 615.
 Jennings, W. H. & Bixler, M. E. 1934 *J. Phys. Chem.* **38**, 747.
 Kistiakowsky, G. B. & Rice, W. W. 1939 *J. Chem. Phys.* **7**, 281.
 Kistiakowsky, G. B. & Rice, W. W. 1940 *J. Chem. Phys.* **8**, 610.
 Lambert, J. D., Roberts, G. A. H., Rowlinson, J. S. & Wilkinson, V. J. 1949 *Proc. Roy. Soc. A*, **196**, 113.
 Lambert, J. D., Staines, E. N. & Woods, S. D. 1950 *Proc. Roy. Soc. A*, **200**, 262.
 Montgomery, J. B. & de Vries, T. 1942 *J. Amer. Chem. Soc.* **64**, 2375.
 Pitzer, K. S. & Scott, D. W. 1943 *J. Amer. Chem. Soc.* **65**, 803.
 Pitzer, K. S. & Woltner, W. 1940 *J. Amer. Chem. Soc.* **71**, 2842.
 Richards, W. T. & Reid, J. A. 1934 *J. Chem. Phys.* **2**, 206.
 Rowlinson, J. S. 1948 *Nature*, **162**, 820.
 Rowlinson, J. S. 1949 *Trans. Faraday Soc.* **45**, 974.
 Schäfer, K. 1940 *Z. phys. Chem. B*, **46**, 212.
 Stewart, J. L. 1946 *Rev. Sci. Instrum.* **17**, 59.
 Telfair, D. 1942 *J. Chem. Phys.* **10**, 167.
 Vold, R. D. 1935 *J. Amer. Chem. Soc.* **57**, 1192.
 Waddington, G. & Deuslin, D. R. 1947 *J. Amer. Chem. Soc.* **69**, 2275.

The invariant theory of isotropic turbulence in magneto-hydrodynamics

By S. CHANDRASEKHAR, F.R.S., *Yerkes Observatory*

(Received 14 June 1950)

In this paper the invariant theory of isotropic turbulence in magneto-hydrodynamics is developed on the basis of the equations of motion recently derived by Batchelor to describe the hydrodynamics of an incompressible fluid which is also a good electrical conductor. The theory allows for the interaction between the electromagnetic field and the turbulent motion when there is no externally imposed electric or magnetic field.

Various double and triple correlation tensors involving the components of the velocity and the magnetic field intensity are defined, and these equations governing the scalars defining these tensors are derived; these latter equations admit integrals of Lofgrensky's type. The equations governing the dissipation of energy by viscosity and conductivity are also derived; they exhibit the manner in which energy is exchanged between the velocity and the magnetic fields. Finally, the equations appropriate for the case, when an external agency supplies kinetic energy to the system at a constant rate and a stationary condition prevails, are obtained; they suggest that the energy in the magnetic field is contained, principally, in the eddies with large wave numbers.

1. INTRODUCTION

In a recent paper Batchelor (1950) has made an important contribution to the systematic study of the interaction between the electromagnetic and the hydrodynamic forces in an incompressible conducting fluid in turbulent motion. In this paper the study initiated by Batchelor will be continued somewhat further. It will be shown in particular that the phenomenological theory of turbulence in magneto-hydrodynamics can be developed to the same extent that the corresponding theory in the absence of electromagnetic forces has been developed by Taylor (1935, 1938) von Kármán & Howarth (1938), Robertson (1940), Kolmogoroff (1941*a, b*) and Batchelor (1947). Indeed, it will appear that the inclusion of electromagnetic forces does not introduce any essential difficulty which is not already present in our understanding of ordinary turbulence.

2. THE EQUATIONS OF MAGNETO-HYDRODYNAMICS

When the conductivity, σ , of a fluid tends to infinity, the electric field strength, $\mathbf{E}$, at each point must tend to the value $-\mu\mathbf{u} \times \mathbf{H}/c$, where $\mathbf{u}$ is the velocity and $\mathbf{H}$ is the magnetic field strength at the point considered. μ is the magnetic permeability and c is the velocity of light; otherwise the current

$$\mathbf{j} = \sigma(c\mathbf{E} + \mu\mathbf{u} \times \mathbf{H}) + \rho_e\mathbf{u}/c \quad (1)$$

will become very large even when the slightest mass motions are present. (In (1) ρ_e denotes the excess charge density, so that the term $\rho_e\mathbf{u}/c$ represents the convection current.) Hence, when σ is large we may assume that

$$\mathbf{E} \simeq -\mu\mathbf{u} \times \mathbf{H}/c, \quad (2)$$

a relation which will be increasingly valid as $\sigma \rightarrow \infty$.

An important consequence of the relation (2) is that under the circumstances in which this is a good approximation the energy in the electric field is of the order of $|\mathbf{u}|^2/c^2$ of the energy in the magnetic field and can, therefore, be neglected. Consequently in this approximation (which we shall call the approximation of magneto-hydrodynamics) we have to consider only the interaction between the two fields, $\mathbf{u}$ and $\mathbf{H}$.

In the magneto-hydrodynamic approximation, Maxwell's equation

$$\frac{\kappa}{c} \frac{\partial \mathbf{E}}{\partial t} = \text{curl } \mathbf{H} - 4\pi \mathbf{j}, \quad (3)$$

where κ is the dielectric constant, becomes

$$\mathbf{j} \simeq \frac{1}{4\pi} \text{curl } \mathbf{H}. \quad (4)$$

In the framework of the approximations (2) and (3), the Stokes-Navier equation modified to take into account the electromagnetic body forces, namely,

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \text{grad}) \mathbf{u} = \frac{1}{\rho} (\rho_e \mathbf{E} + \mu \mathbf{j} \times \mathbf{H}) - \frac{1}{\rho} \text{grad } p + \nu \nabla^2 \mathbf{u}, \quad (5)$$

becomes
$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \text{grad}) \mathbf{u} = \frac{\mu}{4\pi\rho} \text{curl } \mathbf{H} \times \mathbf{H} - \frac{1}{\rho} \text{grad } p + \nu \nabla^2 \mathbf{u}. \quad (6)$$

In equations (5) and (6) p denotes the pressure, ρ the density and ν the kinematic viscosity.

Again, in the approximation (2), Maxwell's equation

$$\frac{\mu}{c} \frac{\partial \mathbf{H}}{\partial t} = -\text{curl } \mathbf{E} \quad (7)$$

becomes
$$\frac{\partial \mathbf{H}}{\partial t} - \text{curl } (\mathbf{u} \times \mathbf{H}) = 0. \quad (8)$$

In a higher approximation in which the loss of energy by Joule heat is allowed for, equation (8) is modified to (cf. Batchelor 1950, equation (2.7))

$$\frac{\partial \mathbf{H}}{\partial t} - \text{curl } (\mathbf{u} \times \mathbf{H}) = \lambda \nabla^2 \mathbf{H}, \quad (9)$$

where
$$\lambda = \frac{1}{4\pi\mu\sigma}. \quad (10)$$

Now the magnetic field intensity $\mathbf{H}$ is a solenoidal vector; and in an incompressible fluid the velocity $\mathbf{u}$ is also a solenoidal vector. When use is made of this property of $\mathbf{H}$ and $\mathbf{u}$, equations (6) and (9) can be written in the forms

$$\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_k} u_i u_k - \frac{\mu}{4\pi\rho} \frac{\partial}{\partial x_k} H_i H_k = -\frac{1}{\rho} \frac{\partial}{\partial x_i} \left(p + \mu \frac{|\mathbf{H}|^2}{8\pi} \right) + \nu \nabla^2 u_i \quad (11)$$

and
$$\frac{\partial H_i}{\partial t} + \frac{\partial}{\partial x_k} (H_i u_k - u_i H_k) = \lambda \nabla^2 H_i, \quad (12)$$

where here and in the sequel summation over the repeated indices is to be understood.

Equations (11) and (12) form the basis of Batchelor's discussion.

In our further work it will be convenient to use instead of $\mathbf{H}$ the quantity

$$\mathbf{h} = \left(\frac{\mu}{4\pi\rho} \right)^{\frac{1}{2}} \mathbf{H}, \quad (13)$$

which is of the dimensions of a velocity. In terms of $\mathbf{h}$ equations (11) and (12) become

$$\frac{\partial u_i}{\partial t} + \frac{\partial}{\partial x_k} (u_i u_k - h_i h_k) = -\frac{1}{\rho} \frac{\partial}{\partial x_i} (p + \frac{1}{2} \rho |\mathbf{h}|^2) + \nu \nabla^2 u_i, \quad (14)$$

and

$$\frac{\partial h_i}{\partial t} + \frac{\partial}{\partial x_k} (h_i u_k - u_i h_k) = \lambda \nabla^2 h_i. \quad (15)$$

3. THE VARIOUS DOUBLE AND TRIPLE CORRELATIONS

It is evident that by introducing various double and triple correlations we can treat equations (14) and (15) in a manner analogous to von Kármán & Howarth's treatment (1938) of the Stokes-Navier equation.

For the sake of simplicity we shall restrict ourselves in this paper to the case of *homogeneous isotropic turbulence*; the extension to the case of axisymmetric turbulence is straightforward and can be carried out if necessary (cf. Chandrasekhar 1950).

From an examination of equations (14) and (15) it is seen that the following double and triple correlations should be considered:

Double correlations:

$$\overline{u_i u'_j}, \quad \overline{h_i h'_j} \quad \text{and} \quad \overline{u_i h'_j}. \quad (16)$$

Triple correlations:

$$\left. \begin{aligned} & \overline{u_i u_j u'_k}, \quad \overline{h_i h_j u'_k}, \quad \overline{u_i u_j h'_k}, \quad \overline{h_i h_j h'_k}, \\ & \overline{(h_i u_j - u_i h_j) h'_k} \quad \text{and} \quad \overline{u_i (h'_j u'_k - h'_k u'_j)}. \end{aligned} \right\} \quad (17)$$

In (16) and (17) the primes and the lack of primes distinguish the values of the quantities taken at x_i and x'_i respectively.

Using the theory of invariants (Robertson 1940; Chandrasekhar 1950) we can write down explicit expressions for the various tensors (16) and (17). However, in the present case attention must be paid to the fact that $\mathbf{h}$, unlike $\mathbf{u}$, is a skew vector. Thus correlations in which the components of $\mathbf{h}$ occur an odd number of times are skew tensors in contrast to those in which they occur an even number of times.

The tensors $\overline{u_i u'_j}$ and $\overline{h_i h'_j}$ are clearly symmetrical and solenoidal in their indices. They can therefore be expressed in the forms (Robertson 1940)

$$\left. \begin{aligned} \overline{u_i u'_j} &= \text{curl } Q \epsilon_{ijl} \xi_l = \frac{Q'}{r} \xi_i \xi_j - (rQ' + 2Q) \delta_{ij}, \\ \overline{h_i h'_j} &= \text{curl } H \epsilon_{ijl} \xi_l = \frac{H'}{r} \xi_i \xi_j - (rH' + 2H) \delta_{ij}, \end{aligned} \right\} \quad (18)$$

where $\xi_i = x'_i - x_i$ and Q and H are two even functions of $r = |\boldsymbol{\xi}|$. In equations (18) (and in the sequel) primes attached to scalar functions (such as Q and H) denote

differentiation with respect to r . Also it may be noted that in (18) the curl is taken with respect to the second index j .

In contrast to $\overline{u_i u'_j}$ and $\overline{h_i h'_j}$ the tensor $\overline{u_i h'_j}$ is skew; accordingly, it must be of the form

$$\overline{u_i h'_j} = R \epsilon_{ijl} \xi_l, \quad (19)$$

where the defining scalar R is again an even function of r .

Turning next to the triple correlations (17), we first observe that the isotropic tensors $\overline{u_i u_j u'_k}$ and $\overline{h_i h_j u'_k}$ are symmetrical in the indices i and j and solenoidal in the index k . Accordingly, they can be derived in a gauge invariant way according to (cf. Chandrasekhar 1950, equations (92) and (94))

$$\begin{aligned} \overline{u_i u_j u'_k} &= \text{curl } T (\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{ikl} \xi_l) \\ &= \frac{2}{r} T' \xi_i \xi_j \xi_k - (rT' + 3T) (\xi_i \delta_{jk} + \xi_j \delta_{ik}) + 2T \delta_{ij} \xi_k, \end{aligned} \quad (20)$$

$$\begin{aligned} \text{and} \quad \overline{h_i h_j u'_k} &= \text{curl } S (\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{ikl} \xi_l) \\ &= \frac{2}{r} S' \xi_i \xi_j \xi_k - (rS' + 3S) (\xi_i \delta_{jk} + \xi_j \delta_{ik}) + 2S \delta_{ij} \xi_k, \end{aligned} \quad (21)$$

where the curl is taken with respect to the third index k .

The scalars T and S introduced in the definitions of $\overline{u_i u_j u'_k}$ and $\overline{h_i h_j u'_k}$ have the following simple meanings:

Considering the correlations between u_1^2 (or u_2^2) at $(0, 0, 0)$ and u'_1 at $(r, 0, 0)$ we have

$$\overline{u_1^2 u'_1} = -4rT(r) \quad \text{and} \quad \overline{u_2^2 u'_1} = +2rT(r). \quad (22)$$

Similarly, the correlations between h_1^2 (or h_2^2) at $(0, 0, 0)$ and u'_1 at $(r, 0, 0)$ are given by

$$\overline{h_1^2 u'_1} = -4rS(r) \quad \text{and} \quad \overline{h_2^2 u'_1} = +2rS(r). \quad (23)$$

Near the origin $T(r)$ must have the behaviour

$$T(r) = T_2 r^2 + O(r^4) \quad (r \rightarrow 0); \quad (24)$$

this follows from the fact that as $r \rightarrow 0$, $\overline{u_1^2 u'_1}$ must behave like r^3 (cf. von Kármán & Howarth 1938). On the other hand, the scalar S must tend to a finite limit S_0 as $r \rightarrow 0$; thus

$$S(r) = S_0 + S_2 r^2 + O(r^4) \quad (r \rightarrow 0). \quad (25)$$

Indeed, from a consideration of the relations (23) for $r \rightarrow 0$ we readily find that

$$S_0 = -\frac{1}{4} \overline{h_1^2 \frac{\partial u_1}{\partial x_1}} = \frac{1}{2} \overline{h_2^2 \frac{\partial u_1}{\partial x_1}}; \quad (26)$$

S_0 is therefore a measure of the contribution to the magnetic energy by the stretching of the magnetic lines of force by the velocity field; for, as we have seen, in the magneto-hydrodynamic approximation the lines of force are 'frozen' with the matter (cf. equation (2)).

The correlations $\overline{u_i u_j h'_k}$ and $\overline{h_i h_j h'_k}$ are skew tensors of the third order, symmetrical in i and j and solenoidal in k . From the symmetry in i and j and the skewness

we can already conclude that (cf. Chandrasekhar 1950, equation (87), where the linearly independent skew tensors of the third order in the general axisymmetric case are listed)

$$\left. \begin{aligned} \overline{u_i u_j h'_k} &= U(\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{ikl} \xi_l) \\ \overline{h_i h_j h'_k} &= V(\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{ikl} \xi_l) \end{aligned} \right\} \quad (27)$$

and

where the defining scalars U and V are even functions of r .

Considering next $\overline{(h_i u_j - h_j u_i) h'_k}$, we observe that this is an isotropic tensor antisymmetrical in i and j . From this it follows that it must be of the form

$$\overline{(h_i u_j - h_j u_i) h'_k} = P(\xi_i \delta_{jk} - \xi_j \delta_{ik}), \quad (28)$$

where the defining scalar P is again an even function of r . It can be readily verified that the tensor (28) is also solenoidal in k as required.

The value of P at the origin is a multiple of S_0 (cf. equation (26)). This can be established in the following manner.

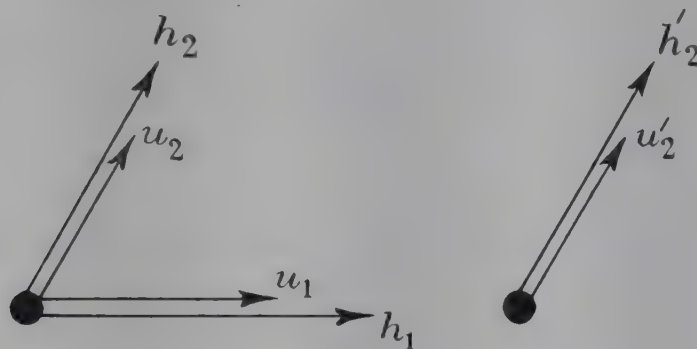


FIGURE 1

According to equation (28) for the configuration of the vector components shown in figure 1, we have

$$\overline{(h_1 u_2 - h_2 u_1) h'_2} = P x_1. \quad (29)$$

Letting $x_1 \rightarrow 0$ in this equation, we obtain

$$\begin{aligned} P_0 &= \overline{(h_1 u_2 - h_2 u_1) \frac{\partial h_2}{\partial x_1}} = -h_2 \frac{\partial}{\partial x_1} \overline{h_1 u_2} + \frac{1}{2} h_2^2 \frac{\partial u_1}{\partial x_1} \\ &= -h_1 h_2 \frac{\partial u_2}{\partial x_1} - h_2 u_2 \frac{\partial h_1}{\partial x_1} + \frac{1}{2} h_2^2 \frac{\partial u_1}{\partial x_1}. \end{aligned} \quad (30)$$

Now using the solenoidal property of h_i , we have

$$\begin{aligned} h_2 u_2 \frac{\partial h_1}{\partial x_1} &= -h_2 u_2 \left(\frac{\partial h_2}{\partial x_2} + \frac{\partial h_3}{\partial x_3} \right) \\ &= +\frac{1}{2} h_2^2 \frac{\partial u_2}{\partial x_2} - h_2 u_2 \frac{\partial h_3}{\partial x_3}. \end{aligned} \quad (31)$$

But on the assumption of isotropy

$$h_2 u_2 \frac{\partial h_1}{\partial x_1} = h_2 u_2 \frac{\partial h_3}{\partial x_3} \quad \text{and} \quad h_2^2 \frac{\partial u_2}{\partial x_2} = h_1^2 \frac{\partial u_1}{\partial x_1}. \quad (32)$$

Hence combining equations (30), (31) and (32) we have

$$P_0 = -\overline{h_1 h_2 \frac{\partial u_2}{\partial x_1}} - \frac{1}{4} \overline{h_1^2 \frac{\partial u_1}{\partial x_1}} + \frac{1}{2} \overline{h_2^2 \frac{\partial u_1}{\partial x_1}}, \quad (33)$$

or, using the result expressed by equation (26), we have

$$P_0 = -\overline{h_1 h_2 \frac{\partial u_2}{\partial x_1}} - \frac{1}{2} \overline{h_1^2 \frac{\partial u_1}{\partial x_1}}. \quad (34)$$

On the other hand, according to equation (21), for the disposition of the vector components shown in figure 1, we have

$$\overline{h_1 h_2 u_2'} = -(rS' + 3S)x_1; \quad (35)$$

now passing to the limit $x_1 = 0$ in this equation, we obtain

$$\overline{h_1 h_2 \frac{\partial u_2}{\partial x_1}} = -3S_0 = +\frac{3}{4} \overline{h_1^2 \frac{\partial u_1}{\partial x_1}}. \quad (36)$$

Hence

$$P_0 = -\frac{5}{4} \overline{h_1^2 \frac{\partial u_1}{\partial x_1}} = 5S_0, \quad (37)$$

which is the required relation.

Finally, considering $\overline{u_i(h_j' u_k' - h_k' u_j')}$, we have a skew isotropic tensor solenoidal in i and antisymmetrical in j and k . From these facts it follows from the theory of invariants that

$$\overline{u_i(h_j' u_k' - h_k' u_j')} = 2W\epsilon_{ijk} + \frac{W'}{r}(\xi_j \epsilon_{kil} \xi_l + \xi_k \epsilon_{ijl} \xi_l), \quad (38)$$

where W is a defining scalar. An alternative form of equation (38) is (cf. Chandrasekhar 1950, equation (85))

$$\overline{u_i(h_j' u_k' - h_k' u_j')} = (2W + rW')\epsilon_{ijk} - \frac{W'}{r}\xi_i \epsilon_{jkl} \xi_l. \quad (39)$$

4. THE EQUATIONS GOVERNING THE SCALARS

From equations (14) and (15) we can derive the equations governing the various scalars introduced in the preceding section by following a procedure first used by von Kármán & Howarth (1938). Thus, multiplying equation (14) by the velocity component u_j' at x_i' and averaging the resulting equation, we obtain

$$\overline{u_j' \frac{\partial u_i}{\partial t}} + \frac{\partial}{\partial x_k} \overline{(u_i u_k u_j' - h_i h_k u_j')} = \nu \nabla^2 \overline{u_i u_j'}. \quad (40)^*$$

Interchanging the indices i and j and also the primed and the unprimed quantities in equation (40) we have

$$\overline{u_i \frac{\partial u_j'}{\partial t}} + \frac{\partial}{\partial x_k'} \overline{(u_j' u_k' u_i - h_j' h_k' u_i)} = \nu \nabla^2 \overline{u_i u_j'}. \quad (41)$$

* The term in the pressure in the equations of motion does not appear in this and similar equations since there exists no non-vanishing isotropic solenoidal vector.

Adding equations (40) and (41) and remembering that

$$\frac{\partial}{\partial \xi_i} = \frac{\partial}{\partial x'_i} = -\frac{\partial}{\partial x_i}, \quad (42)$$

we have

$$\frac{\partial}{\partial t} \overline{u_i u'_j} - 2 \frac{\partial}{\partial \xi_k} \overline{(u_i u_k u'_j - h_i h_k u'_j)} = 2\nu \nabla^2 \overline{u_i u'_j}. \quad (43)$$

In deriving this equation we have made use of the relations

$$\overline{u_i u_j u'_k} = -\overline{u'_i u'_j u_k} \quad \text{and} \quad \overline{h_i h_j u'_k} = -\overline{h'_i h'_j u_k}. \quad (44)$$

Now according to equation (20)

$$\begin{aligned} \frac{\partial}{\partial \xi_k} \overline{u_i u_k u'_j} &= \text{curl} \frac{\partial}{\partial \xi_k} T(\xi_i \epsilon_{kjl} \xi_l + \xi_k \epsilon_{ijl} \xi_l) \\ &= \text{curl} (rT' + 5T) \epsilon_{ijl} \xi_l. \end{aligned} \quad (45)$$

Similarly,

$$\frac{\partial}{\partial \xi_k} \overline{h_i h_k u'_j} = \text{curl} (rS' + 5S) \epsilon_{ijl} \xi_l. \quad (46)$$

Further, it is known that the scalar defining the Laplacian of a second-order isotropic solenoidal vector is obtained by operating

$$D = \left(\frac{\partial^2}{\partial r^2} + \frac{4}{r} \frac{\partial}{\partial r} \right) \quad (47)$$

on the scalar defining the original tensor (cf. Robertson 1940). The scalars defining the various second-order tensors occurring in equation (43) are, therefore,

$$\frac{\partial Q}{\partial t}, \quad \left(r \frac{\partial}{\partial r} + 5 \right) (T - S) \quad \text{and} \quad D(Q), \quad (48)$$

respectively. Hence equation (43) is equivalent to

$$\frac{\partial Q}{\partial t} = 2 \left(r \frac{\partial}{\partial r} + 5 \right) (T - S) + 2\nu D(Q). \quad (49)$$

This is the generalization of the usual von Kármán-Howarth equation to magneto-hydrodynamics.

Equation (15) can be treated in an analogous manner by multiplying by h'_j , averaging the resulting equation, etc. In this manner we obtain

$$\frac{\partial}{\partial t} \overline{h_i h'_j} - 2 \frac{\partial}{\partial \xi_k} \overline{(h_i u_k - h_k u_i) h'_j} = 2\lambda \nabla^2 \overline{h_i h'_j}. \quad (50)$$

From equation (28) defining the triple correlation occurring in equation (50), we find

$$\begin{aligned} \frac{\partial}{\partial \xi_k} \overline{(h_i u_k - h_k u_i) h'_j} &= \frac{\partial}{\partial \xi_k} P(\xi_i \delta_{kj} - \xi_k \delta_{ij}) \\ &= \frac{P'}{r} \xi_i \xi_j - (rP' + 2P) \delta_{ij} = \text{curl} P \epsilon_{ijl} \xi_l. \end{aligned} \quad (51)$$

The scalars defining the three second-order tensors occurring in equation (50) are therefore

$$\frac{\partial H}{\partial t}, \quad P \quad \text{and} \quad D(H). \quad (52)$$

Hence equation (50) is equivalent to

$$\frac{\partial H}{\partial t} = 2P + 2\lambda D(H). \quad (53)$$

A third equation governing the scalars can be obtained in the following manner: Multiply equation (14) by h'_j and average. We obtain

$$\overline{h'_j \frac{\partial u_i}{\partial t}} - \frac{\partial}{\partial \xi_k} \overline{u_i u_k h'_j} + \frac{\partial}{\partial \xi_k} \overline{h_i h_k h'_j} = \nu \nabla^2 \overline{u_i h'_j}. \quad (54)$$

Next, consider equation (15) in the primed variables for h'_j , multiply this equation by u_i and average. We obtain

$$\overline{u_i \frac{\partial h'_j}{\partial t}} + \frac{\partial}{\partial \xi_k} \overline{(h'_j u'_k - u'_j h'_k) u_i} = \lambda \nabla^2 \overline{u_i h'_j}. \quad (55)$$

Finally, adding equations (54) and (55) we have

$$\frac{\partial}{\partial t} \overline{u_i h'_j} - \frac{\partial}{\partial \xi_k} \overline{(u_i u_k h'_j - h_i h_k h'_j)} + \frac{\partial}{\partial \xi_k} \overline{u_i (h'_j u'_k - u'_j h'_k)} = (\lambda + \nu) \nabla^2 \overline{u_i h'_j}. \quad (56)$$

Now according to equations (27) and (38)

$$\left. \begin{aligned} \frac{\partial}{\partial \xi_k} \overline{(u_i u_k h'_j - h_i h_k h'_j)} &= \left(r \frac{\partial}{\partial r} + 5 \right) (U - V) \epsilon_{ijl} \xi_l \\ \text{and} \quad \frac{\partial}{\partial \xi_k} \overline{u_i (h'_j u'_k - u'_j h'_k)} &= D(W) \epsilon_{ijl} \xi_l \end{aligned} \right\} \quad (57)$$

Hence equation (56) is equivalent to

$$\frac{\partial R}{\partial t} = \left(r \frac{\partial}{\partial r} + 5 \right) (U - V) - D(W) + (\lambda + \nu) D(R). \quad (58)$$

Equations (49), (53) and (58) are the basic equations of our problem.

5. THE EQUATIONS IN TERMS OF THE VECTOR POTENTIAL

It is sometimes convenient to introduce the vector potential $\mathbf{A}$ according to its definition

$$\mathbf{H} = \text{curl } \mathbf{A}. \quad (59)$$

Under the conditions of our problem, no loss of generality will be entailed by the assumption that $\mathbf{A}$ is solenoidal.

Analogous to $\mathbf{h}$ (cf. equation (13)) define

$$\mathbf{a} = \left(\frac{\mu}{4\pi\rho} \right)^{\frac{1}{2}} \mathbf{A}, \quad (60)$$

so that

$$\mathbf{h} = \text{curl } \mathbf{a}. \quad (61)$$

In terms of the components of $\mathbf{a}$ we shall now define the correlations (cf. (16) and (17))

$$\overline{a'_i a'_j}, \quad \overline{u'_i a'_j}, \quad \overline{u'_i u'_j a'_k}, \quad \overline{h'_i h'_j a'_k} \quad \text{and} \quad \overline{(h'_i u'_j - h'_j u'_i) a'_k}. \quad (62)$$

In view of the relation (61) between $\mathbf{h}$ and $\mathbf{a}$ the scalars defining the tensors (62) are simply related to the scalars defining the corresponding tensors in which the components of $\mathbf{h}$ are substituted where those of $\mathbf{a}$ occur. Thus if A and B are the defining scalars of the isotropic solenoidal tensors $\overline{a_i a_j}$ and $\overline{u_i a_j}$, then it can be readily shown that

$$H = -D(A) \quad \text{and} \quad R = -D(B), \quad (63)$$

where D is the differential operator (47). Similarly, considering $\overline{u_i u_j a'_k}$ we observe that this is an isotropic tensor symmetrical in i and j and solenoidal in k ; it must therefore be expressible in the form (cf. equations (20) and (21))

$$\overline{u_i u_j a'_k} = \text{curl } E(\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{ikl} \xi_l). \quad (64)$$

But $\overline{u_i u_j h'_k}$ is the curl of $\overline{u_i u_j a'_k}$ with respect to k ; and since the tensor on which the curl operates in (64) is already solenoidal in k , it is evident that

$$\overline{u_i u_j h'_k} = -\nabla^2 E(\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{ikl} \xi_l). \quad (65)$$

The Laplacian on the right-hand side can be evaluated and leads to the result

$$\overline{u_i u_j h'_k} = -\left(\frac{\partial^2 E}{\partial r^2} + \frac{6}{r} \frac{\partial E}{\partial r}\right)(\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{ikl} \xi_l), \quad (66)$$

and comparing this with the earlier representation (28), we conclude that

$$U = -\left(\frac{\partial^2 E}{\partial r^2} + \frac{6}{r} \frac{\partial E}{\partial r}\right). \quad (67)$$

$$\text{Similarly} \quad V = -\left(\frac{\partial^2 F}{\partial r^2} + \frac{6}{r} \frac{\partial F}{\partial r}\right), \quad (68)$$

if F is the defining scalar of $\overline{h_i h_j a'_k}$.

Finally, considering the last of the tensors (62), we must have (cf. equation (38))

$$(\overline{h_i u_j} - \overline{h_j u_i}) a'_k = 2G \epsilon_{ijk} + \frac{G'}{r} (\xi_i \epsilon_{jkl} \xi_l + \xi_j \epsilon_{kil} \xi_l), \quad (69)$$

where G is an even function of r ; this representation follows from the fact that the tensor on the left-hand side is skew, antisymmetrical in i and j and solenoidal in k . But

$$(\overline{h_i u_j} - \overline{h_j u_i}) h'_k = \epsilon_{klm} \frac{\partial}{\partial \xi_l} (\overline{h_i u_j} - \overline{h_j u_i}) a'_m. \quad (70)$$

Evaluating the quantity on the right-hand side according to equation (69) and comparing it with the representation (28), we obtain the relation

$$P = -D(G). \quad (71)$$

The equations governing the scalars A , B , E , F and G can be found quite readily from equations (53) and (58). Thus combining equations (53), (63) and (71), we have

$$D\left(\frac{\partial A}{\partial t}\right) = 2\lambda D^2(A) + 2D(G). \quad (72)$$

Now if a function $f(r)$ satisfies the differential equation

$$D(f) = \frac{\partial^2 f}{\partial r^2} + \frac{4}{r} \frac{\partial f}{\partial r} = 0, \quad (73)$$

then

$$r^4 f' = \text{constant.} \quad (74)$$

The continuity of f at $r = 0$ therefore requires that

$$f' = 0 \quad \text{or} \quad f = \text{constant.} \quad (75)$$

Hence from equation (72) it follows that

$$\frac{\partial A}{\partial t} = 2\lambda D(A) + 2G + \text{constant.} \quad (76)$$

Similarly, combining equations (58), (63), (67) and (68) we have

$$D\left(\frac{\partial B}{\partial t}\right) = \left(r\frac{\partial}{\partial r} + 5\right)\left(\frac{\partial^2}{\partial r^2} + \frac{6}{r}\frac{\partial}{\partial r}\right)(E - F) + D(W) + (\lambda + \nu) D^2(B). \quad (77)$$

$$\text{Using the identity} \quad D\left(r\frac{\partial}{\partial r} + 5\right) \equiv \left(r\frac{\partial}{\partial r} + 5\right)\left(\frac{\partial^2}{\partial r^2} + \frac{6}{r}\frac{\partial}{\partial r}\right), \quad (78)$$

we find that equation (77) is equivalent to

$$\frac{\partial B}{\partial t} = \left(r\frac{\partial}{\partial r} + 5\right)(E - F) + W + (\lambda + \nu) D(B) + \text{constant.} \quad (79)$$

6. THE DISSIPATION OF ENERGY BY VISCOSITY AND CONDUCTIVITY

The equations governing the rate of dissipation of energy can be derived from equations (49) and (53) in the following manner:

Let (cf. equations (24) and (25))

$$\left. \begin{aligned} Q &= Q_0 + Q_2 r^2 + \dots, & H &= H_0 + H_2 r^2 + \dots, \\ S &= S_0 + S_2 r^2 + \dots, & \text{and} \quad T &= T_2 r^2 + \dots, \end{aligned} \right\} \quad (80)$$

represent the series expansions for the various scalars valid near $r = 0$. Substituting these expansions in equations (49) and (53) and equating the coefficients of the terms independent of r , we obtain

$$\frac{dQ_0}{dt} = -10S_0 + 20\nu Q_2 \quad \text{and} \quad \frac{dH_0}{dt} = 2P_0 + 20\lambda H_2. \quad (81)$$

On the other hand, according to equations (18)

$$\left. \begin{aligned} \overline{u^2} &= [\overline{u_i u'_i}]_{r=0} = -6Q_0 \\ \overline{h^2} &= [\overline{h_i h'_i}]_{r=0} = -6H_0 \end{aligned} \right\} \quad (82)$$

and

$$\text{Also (cf. equation (37))} \quad P_0 = 5S_0 = -\frac{5}{4} h_1^2 \frac{\partial u_1}{\partial x_1}. \quad (83)$$

Using these relations in equations (81), we obtain

$$\frac{1}{2} \frac{d\overline{u^2}}{dt} = -7.5 h_1^2 \frac{\partial u_1}{\partial x_1} - 60\nu Q_2$$

and

$$\frac{1}{2} \frac{d\overline{h^2}}{dt} = +7.5 h_1^2 \frac{\partial u_1}{\partial x_1} - 60\lambda H_2. \quad (84)$$

These equations give the rates of change of kinetic and the magnetic energy in the system per unit mass. Moreover, according to these equations the rate of change of the total energy is given by

$$\frac{1}{2} \frac{d}{dt} (\overline{u^2} + \overline{h^2}) = -60(\nu Q_2 + \lambda H_2). \quad (85)$$

The physical meaning of this equation is that the rate of loss of energy from the system is entirely due to dissipation either by viscosity in the form of molecular motion or by conductivity in the form of Joule heat. Also, the appearance of the term $7.5 \overline{h_1^2 \partial u_1 / \partial x_1}$ in equations (84) with opposite signs means that the loss in kinetic energy due to the stretching of the magnetic lines of force is compensated by the gain of an equal amount of energy in the magnetic field.

7. INVARIANTS OF LOITSIANSKY'S TYPE FOR MAGNETO-HYDRODYNAMICS

It is known (Loitsiansky 1939; Batchelor & Townsend 1948; and Batchelor 1949) that in ordinary turbulence the integral

$$\Lambda = \int_0^\infty Q r^4 dr, \quad (86)$$

where Q is the defining scalar of the fundamental correlation tensor $\overline{u_i u_j'}$ exists and is constant during the process of the decay of turbulence. As Batchelor (1949) has shown the meaning of this constancy of Λ is that 'the spectral density at very low wave numbers is permanent and is determined by the initial conditions'. It will now be shown that the equations of magneto-hydrodynamics also allow invariants of this type.

Multiplying equation (49) by r^4 and integrating from 0 to r we obtain

$$\frac{\partial}{\partial t} \int_0^r Q r^4 dr = 2r^5(T - S) + 2\nu r^4 \frac{\partial Q}{\partial r}. \quad (87)$$

$$\text{Consequently, if } r^5(T - S) \rightarrow 0 \text{ and } r^4 Q' \rightarrow 0 \text{ as } r \rightarrow \infty, \quad (88)$$

$$\text{then } \int_0^\infty Q r^4 dr = \text{constant}. \quad (89)$$

Assumptions analogous to (88) underlie the usual derivation of Loitsiansky's invariant from the von Kármán-Howarth equation. Perhaps it is not unreasonable to expect that similar conditions will also prevail in magneto-hydrodynamics, in which case we shall have a Loitsiansky's invariant exactly as in ordinary turbulence; and the meaning of this invariant would again be the same as in ordinary turbulence.

Equation (58) also admits an integral under similar conditions: For, multiplying this equation by r^4 and integrating from 0 to r , we obtain

$$\frac{\partial}{\partial t} \int_0^r R r^4 dr = r^5(U - V) - r^4 \frac{\partial W}{\partial r} + (\lambda + \nu) r^4 \frac{\partial R}{\partial r}. \quad (90)$$

Hence, if $r^5(U - V) \rightarrow 0$, $r^4 W' \rightarrow 0$ and $r^4 R' \rightarrow 0$ as $r \rightarrow \infty$, (91)

then
$$\int_0^\infty R r^4 dr = \text{constant.} \quad (92)$$

As it stands, equation (53) for H does not admit an invariant of Loitsiansky's type. However, if we substitute for P according to equation (71), then

$$\frac{\partial}{\partial t} \int_0^r r^4 H dr = -2r^4 \frac{\partial G}{\partial r} + 2\lambda r^4 \frac{\partial H}{\partial r}. \quad (93)$$

Hence, if $r^4 G' \rightarrow 0$ and $r^4 H' \rightarrow 0$ as $r \rightarrow \infty$, (94)

then
$$\int_0^\infty H r^4 dr = \text{constant.} \quad (95)$$

On the other hand, if we are permitted to substitute for P according to equation (71), it may also be permissible to substitute for H according to equation (63) in which case

$$\int_0^r r^4 H dr = r^4 \frac{\partial A}{\partial r}, \quad (96)$$

and we might be tempted to conclude that

$$\int_0^\infty r^4 H dr = 0 \quad (r^4 A' \rightarrow 0 \text{ as } r \rightarrow \infty). \quad (97)$$

In other words, it would appear that if a Loitsiansky invariant for H exists, then it is probably zero. From one point of view this last conclusion may indeed be expected and a correct one; for, as Batchelor (1950) has emphasized, there exists a close analogy between the magnetic field $\mathbf{H}$ and the vorticity $\boldsymbol{\omega}$, and it is known that

$$\int_0^\infty \Omega r^4 dr = 0, \quad (98)$$

where Ω is the defining scalar of $\overline{\omega_i \omega'_j}$.

Finally, it should be noted that under the same conditions as in the foregoing paragraph it would follow from the expression for R given in equation (63) that

$$\int_0^\infty R r^4 dr = 0 \quad (r^4 B' \rightarrow 0 \text{ as } r \rightarrow \infty). \quad (99)$$

8. THE EQUATIONS GOVERNING STATIONARY TURBULENCE IN MAGNETO-HYDRODYNAMICS

Just as in the case of ordinary turbulence, we can conceive in magneto-hydrodynamics, also, of a state of affairs in which an external agency supplies energy to the system at a constant rate and a stationary condition prevails. However, in the present case, since there is no imposed electric or magnetic field it will be natural to suppose that the external agency supplies energy to the system only in the form of kinetic energy and the magnetic energy which exists has arisen spontaneously in the manner envisaged, for example, by Batchelor (1950).

Let ϵ , then, denote the rate at which kinetic energy is being supplied to the system per unit mass. Under stationary conditions ϵ must also be the rate at which energy is dissipated by viscosity and in the process of stretching the magnetic lines of force. The energy transformed into magnetic energy by this latter process of the stretching of the lines of force must in turn be dissipated in the form of Joule heat. According to equations (84), we can express these conditions by

$$\left. \begin{aligned} \epsilon &= +7.5 h_1^2 \frac{\partial u_1}{\partial x_1} + 60 \nu Q_2 \\ 0 &= -7.5 h_1^2 \frac{\partial u_1}{\partial x_1} + 60 \lambda H_2, \end{aligned} \right\} \quad (100)$$

and

where Q_2 and H_2 are now considered constants independent of time.

From equations (100) it follows that

$$\epsilon = 60(\nu Q_2 + \lambda H_2); \quad (101)$$

this equation merely states that in a steady state the system dissipates energy by viscosity and conductivity at the rate at which energy is supplied to the system.

It may be noted that the conditions expressed by equations (100) are equivalent to those under which Batchelor has discussed this problem.

Equations (100) imply that under stationary conditions equations (49) and (53) must be replaced by

$$\frac{1}{6}\epsilon = \left(r \frac{d}{dr} + 5\right)(T - S) + \frac{\nu}{r^4} \frac{d}{dr} \left(r^4 \frac{dQ}{dr}\right) \quad (102)$$

and

$$0 = P + \frac{\lambda}{r^4} \frac{d}{dr} \left(r^4 \frac{dH}{dr}\right), \quad (103)$$

where T , S , Q , P and H are now as functions of r only; for, if the series expansions (80) are substituted in these equations we recover equations (100) as those governing the dominant terms.

By multiplying equation (102) by r^4 and integrating from 0 to r we obtain

$$\frac{1}{30}\epsilon r = r(T - S) + \nu \frac{dQ}{dr}. \quad (104)$$

Equation (103) can also be integrated if we substitute for P according to equation (71); for, with this latter substitution equation (103) becomes

$$0 = \frac{1}{r^4} \frac{d}{dr} \left(r^4 \frac{d}{dr} [\lambda H - G]\right). \quad (105)$$

Hence (cf. equations (73) to (75))

$$\lambda H - G = \text{constant}. \quad (106)$$

Equation (106) is of exactly the same form as the one satisfied by the defining scalar of $\overline{\omega_i \omega_j'}$ in the theory of ordinary turbulence.

It is not immediately clear whether under stationary conditions $\partial R/\partial t$ in equation (58) should be set equal to zero or a constant. It would appear that this depends on

whether the Loitsiansky invariant for R is zero or not. However, we may expect that by setting

$$\frac{\partial R}{\partial t} = \frac{1}{3}\zeta = \text{constant}, \quad (107)$$

we shall not lose any generality. On this assumption equation (58) becomes

$$\frac{1}{3}\zeta = \left(r \frac{d}{dr} + 5\right)(U - V) + \frac{1}{r^4} \frac{d}{dr} \left(r^4 \frac{d}{dr} [(\lambda + \nu)R - W]\right). \quad (108)$$

This equation can be integrated once to give

$$\frac{1}{15}\zeta r = r(U - V) + (\lambda + \nu) \frac{dR}{dr} - \frac{dW}{dr}. \quad (109)$$

Equations (104), (106) and (109) may now be used to discuss the behaviour of the turbulence for large Reynolds numbers in the manner of Kolmogoroff's theory of locally isotropic turbulence (cf. Batchelor 1947). Thus, from equation (104), we may conclude that as $r \rightarrow \infty$

$$(T - S) \rightarrow \frac{\epsilon}{30}. \quad (110)$$

According to equations (22) and (23) we may rewrite this relation in the form

$$\overline{(u_1^2 + 2h_2^2)} u_1' \rightarrow -\frac{2}{15} \epsilon r. \quad (111)$$

This is the analogue of a similar relation in ordinary turbulence (cf. Batchelor 1947, equation (5.5)).

Similarly, from equation (109) we may conclude that

$$(U - V) \rightarrow \frac{1}{15}\zeta. \quad (112)$$

If the Loitsiansky invariant for R should be zero, then

$$U - V \rightarrow 0 \quad (r \rightarrow \infty). \quad (113)$$

The relations (106), (111) and particularly (113) suggest that under stationary conditions the energy in the magnetic field is contained in wave numbers which are small compared to those in which the kinetic energy is contained; they, therefore, confirm in a general way Batchelor's ideas on the subject.

In order to avoid misunderstanding it should be stated that the analysis in this section strictly applies only to small values of r ; for, the assumption of an external agency supplying energy in the form of kinetic energy is strictly incompatible with the assumptions of homogeneity and isotropy. However, it is still possible, as in Kolmogoroff's theory, to discuss the motion of the small eddies as effectively homogeneous and isotropic; it is to these small eddies that the discussion in this section applies. In particular, in setting the left-hand side of equation (103) equal to zero under these circumstances, we are effectively making the same assumption as Batchelor, that the magnetic energy is chiefly associated with large wave numbers. On this last assumption ζ will be zero since R is also determined by the magnetic field. Finally, the validity of equations (110) to (113) requires that r is small but $r \gg (\nu^3/\epsilon)^{\frac{1}{4}}$.

In conclusion I wish to record my indebtedness to Dr G. K. Batchelor for allowing me to read a manuscript copy of his paper in advance of its publication.

REFERENCES

- Batchelor, G. K. 1947 *Proc. Camb. Phil. Soc.* **43**, 533.
 Batchelor, G. K. 1949 *Proc. Roy. Soc. A*, **195**, 513.
 Batchelor, G. K. 1950 *Proc. Roy. Soc. A*, **201**, 405.
 Batchelor, G. K. & Townsend, A. A. 1948 *Proc. Roy. Soc. A*, **194**, 527.
 Chandrasekhar, S. 1950 *Phil. Trans. A*, **242**, 557.
 Kármán, T. von & Howarth, L. 1938 *Proc. Roy. Soc. A*, **164**, 192.
 Kolmogoroff, A. N. 1941a *C.R. Acad. Sci. U.R.S.S.* **30**, 301.
 Kolmogoroff, A. N. 1941b *C.R. Acad. Sci. U.R.S.S.* **32**, 16.
 Loitsiansky, L. 1939 *Cent. Aero. Hydro. Inst. (Moscow), Rep.* 440. (Trans. as *N.A.C.A. Tech. Memo*, no. 1079.)
 Robertson, H. P. 1940 *Proc. Camb. Phil. Soc.* **36**, 209.
 Taylor, G. I. 1935 *Proc. Roy. Soc. A*, **151**, 421.
 Taylor, G. I. 1938 *Proc. Roy. Soc. A*, **164**, 15.

The effect of an electric field on the viscosity of liquids. II

BY E. N. DA C. ANDRADE, F.R.S., AND C. DODD, PH.D.

(Received 2 August 1950)

To obtain the true effect of an electric field on the viscosity of liquids measurements must be made with alternating fields whose frequency lies within a certain range. As the effect is very small a very sensitive method of measurements of viscosity must be devised and various temperature effects must be eliminated. Apparatus and methods are described which allow comparative viscosities to be consistently measured, with and without field, to within a few parts in a million. With this disposition it has been shown that for non-polar benzene there is no effect as large as 1 part in 100,000, even with the highest applied fields. With polar liquids, however, there is a true, but small, increase of viscosity proportional to the square of the field. A brief theoretical discussion is given, in which it is shown that electrostriction cannot be responsible for the effect.

INTRODUCTION

In a previous paper (Andrade & Dodd 1946), referred to subsequently as part I, we recorded experiments on the influence of a constant electric field on the viscosity of liquids. It was shown that, while there was no detectable effect with non-polar liquids, or with polar liquids that were good insulators, a considerable increase in viscosity took place when polar liquids that transmitted a small current were subjected to high fields. This effect was traced to the action of the field in producing concentrations of free ions in the neighbourhood of the metallic surfaces between which the liquid flowed, these ions acting as centres round which the polar molecules formed clusters. On this basis the conflicting results of previous experiments were made clear, and various features of the effect established by us were quantitatively explained. It was shown, for instance, that with alternating fields of low frequencies the full effect existed, but that as the frequency was increased a stage

was reached at which the effect rapidly diminished, and at high frequencies, when the ions had no time to migrate appreciably, the effect was not measurable with the experimental disposition then at our disposal.

At the end of part I we set out the conditions under which a true effect of electric field on liquid viscosity was to be sought, namely, by the use of an alternating field whose frequency was well in excess of uV/D^2 , where u is the ionic mobility, V the applied potential and D the separation of the two electrodes between which the liquid flows. A very sensitive method of measuring small changes of viscosity is also necessary. In the following pages are described experiments under the required conditions which have led to the definite establishment with three polar liquids of a small increase of viscosity proportional to the square of the field. It has also been shown to a much higher degree of precision than hitherto that there is no such effect with the non-polar liquid benzene.

EXPERIMENTAL ARRANGEMENT

(a) *Viscometer*

The viscometer is similar in pattern to that described in part I, the liquid flowing in a rectangular channel between plane steel surfaces, which are separated by insulating plates (see figures 1*a* and 1*b*, part I). These insulating plates are, however, made of optically worked Pyrex glass, instead of the quartz glass used in the earlier experiments, since this reduces the difference of thermal expansion between the insulation and the steel. The depth of the channel is reduced to 0.143 mm., as against a least value of 0.20 mm. in the earlier apparatus, which both increases the time of flow and reduces the potential difference required to produce a given field.

(b) *The alternating field*

The frequency of the alternating voltage supply is in the neighbourhood of 10,000 c./sec., which is well in excess of the greatest value (about 2000) of uV/D^2 , while at the same time well below frequencies at which relaxation effects occur in polar liquids. This supply is produced by a beat-frequency oscillator, whose output is fed into an amplifier capable of delivering 600 V (peak). The leads from the amplifier are connected to the opposite plates of the viscometer by way of 10 mA fuses and an a.c. microammeter. The voltage across the plates is measured by means of a voltmeter placed with the viscometer in the air bath inside the water bath. An external valve voltmeter and cathode-ray tube are also connected in the circuit, to enable the wave form to be checked.

(c) *Temperature control*

For the increased accuracy of measurement contemplated it is necessary to keep the temperature constant within narrow limits. The viscometer is held by a rigid brass framework inside an air-filled box which is totally immersed in the water of a large thermostat. The box (floor 30 × 20 cm., height 40 cm.) is constructed of transparent Perspex sheet, the lid being made liquid-tight by a rubber gasket. The glass-sided thermostat tank contains about 200 l. of water, made to circulate by

stirrers all round the Perspex box, the air within which is also kept in motion by a stirrer. The water tank is further surrounded by a glass-sided air space to minimize the effect of draughts.

The toluene regulator for the water bath consists of thin-walled tubing, about 15 mm. in diameter, wound in a spiral of five turns and containing about a litre of toluene. The capillary in which the intermittent mercury contact operates is so chosen that a change of temperature of the bath of 0.001°C causes a change in level of the mercury in the capillary of about 2 mm. The intermittent water heater is a 40 W carbon filament lamp in series with a rheostat, the current being adjusted so that the heat supplied exceeds the heat lost by a small amount, the on-off periods thus being fairly long. After 24 hr. has been allowed for the system to settle down, no movement on a Beckmann thermometer divided into hundredths of a degree can be observed over a period of many days, so that the temperature can be estimated to be constant within 0.001°C . The time of flow of the liquid in the viscometer, however, furnishes a more pertinent method of measuring temperature changes of the essential part of the apparatus. Times of flow of about 1000 sec., taken at intervals of 4 hr., differ by at most 0.01 sec., and show no systematic trend in either direction, which, with the liquid in question, indicates that the temperature of the viscometer is being maintained constant to better than 0.001°C .

(d) *Control of pressure conditions*

As in the experiments of part I the flow takes place under the difference of level between the liquid standing in tubes *A* and *B* of the viscometer (figure 1), the time

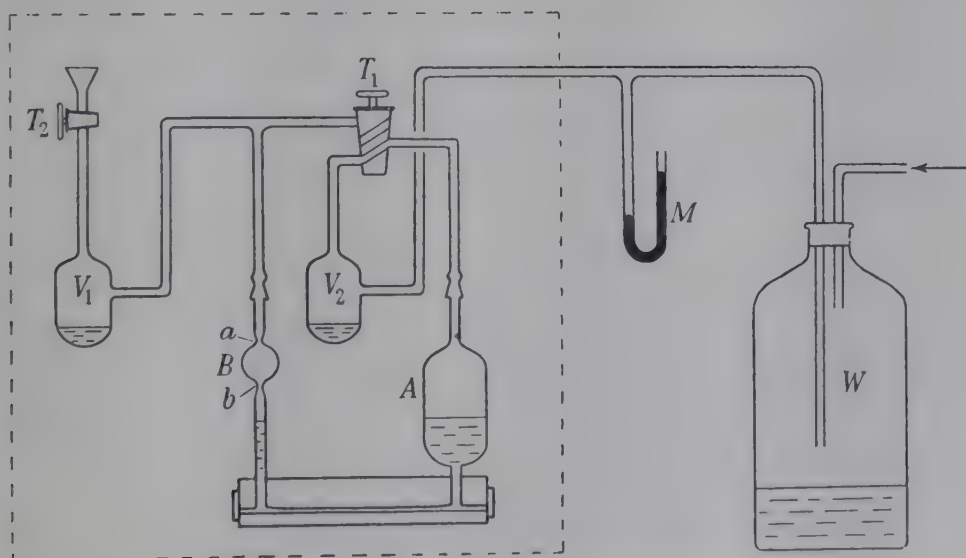


FIGURE 1. General view of apparatus.

measured being that for the liquid in *B* to fall from mark *a* to mark *b*. The difference of level diminishes from about 11 cm. to about 7.5 cm. during an experiment. Since the accuracy depends upon precise reproducibility of the level conditions, it is obviously essential that there shall not be any loss of liquid due to evaporation at any stage of the experiment, including the resetting, during which the liquid is raised in tube *B*.

The procedure is as follows. The viscometer, containing an amount of liquid chosen so as to give a suitable head when set, stands in the thermostat with the level the same in both limbs. The tap T_1 is turned so as to connect the bulb A , called the reservoir bulb, to the large vessel W which contains some of the experimental liquid, the pressure in W being slightly in excess of atmospheric. The tap T_2 is open. When the level of the liquid in B has risen to the required height above a , T_2 is closed and T_1 turned so as to connect A and B and at the same time isolate them from the vessel W . The bulbs V_1 and V_2 (which are, of course, within the thermostatically controlled vessel, indicated by the dotted line) contain each a little of the experimental liquid, and ensure that during all manipulations any air transferred shall contain vapour saturated at the temperature of the experiment.

That this arrangement is satisfactory was shown by the tests (already quoted), made at intervals of 4 hr., which gave times of flow, of duration 1000 sec., consistent within 0.01 sec. and showed no systematic trend with time.

(e) *The timing system*

To achieve the accuracy desired an automatic time-recording system is used. The apparatus* is of a type designed by Dr A. B. Wood and made by Messrs Tinsley, the ultimate control being a standard 25 c./sec. tuning fork which governs a synchronous motor, carrying a drum. The operation of a high-speed mechanical relay brings into contact with this drum a light wheel which governs, through a system of gears, a series of three dials recording seconds, tenths of seconds and hundredths of seconds respectively, it being possible to estimate tenths of divisions and thus read to thousandths of a second. The relay is energized and released by the passage of the liquid meniscus, through the agency of a photoelectric device shortly to be described.

As regards the fork controlling the timing system, the amplitude of vibration, and hence the frequency, varied slightly with the energizing current, which was therefore kept constant. There was also a slight variation of frequency with temperature, of about 1 part in 9000 per °C. The temperature of the fork was measured, but not controlled, and a correction made for the small change due to the slight daily rise in temperature of the room. The timing system was periodically checked against a chronometer, which controlled a mechanism by which electrical contact was made for a standard interval of 24.478 sec. This gave a check on the temperature variation, but the constancy of the timing system throughout a determination, which is what determines the accuracy of the measured variation of viscosity, is 1 or 2 parts in 10^6 .

(f) *The photoelectric control*

The optical system, by which the passage of the meniscus actuates the timer, consists of two units of the type depicted in figure 2.

A diminished image I of a fine, horizontal slit S , illuminated by a 48 W straight filament lamp F , is formed by lens L_1 on the narrowed portion of the vertical tube just above the viscometer bulb (a , figure 1), and this image is in turn projected by a second lens L_2 on to the slit aperture of a photoelectric cell C . To give maximum concentration of light on the cell, the Maxwellian view is employed, an image of the

* Kindly lent to us by the Admiralty.

filament source being formed on the surface of lens L_1 by means of the lens L . A similar illuminating and recording system is used for the narrow portion of the tube below the viscometer bulb (b , figure 1).

Although all the liquids used are transparent, the momentary deviation of the light by the curved meniscus as it passes through the well-defined level of the fall tube where the image I is formed is sufficient to operate momentarily a high speed relay in the amplifier circuit of the photoelectric cell. The positions of all components of the optical system are fixed by the use of rigid optical benches on both sides of the water bath.

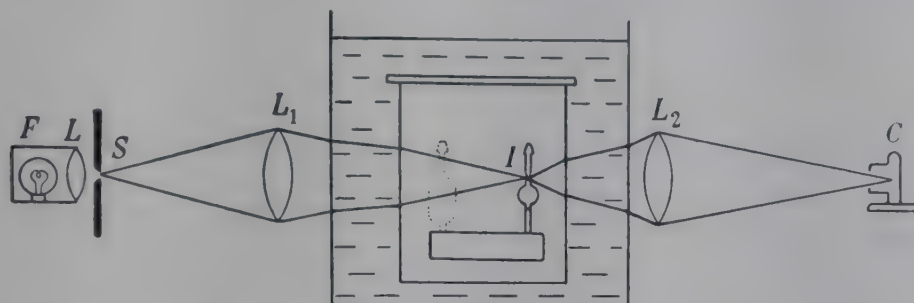


FIGURE 2. Optical system for timing passage of meniscus.

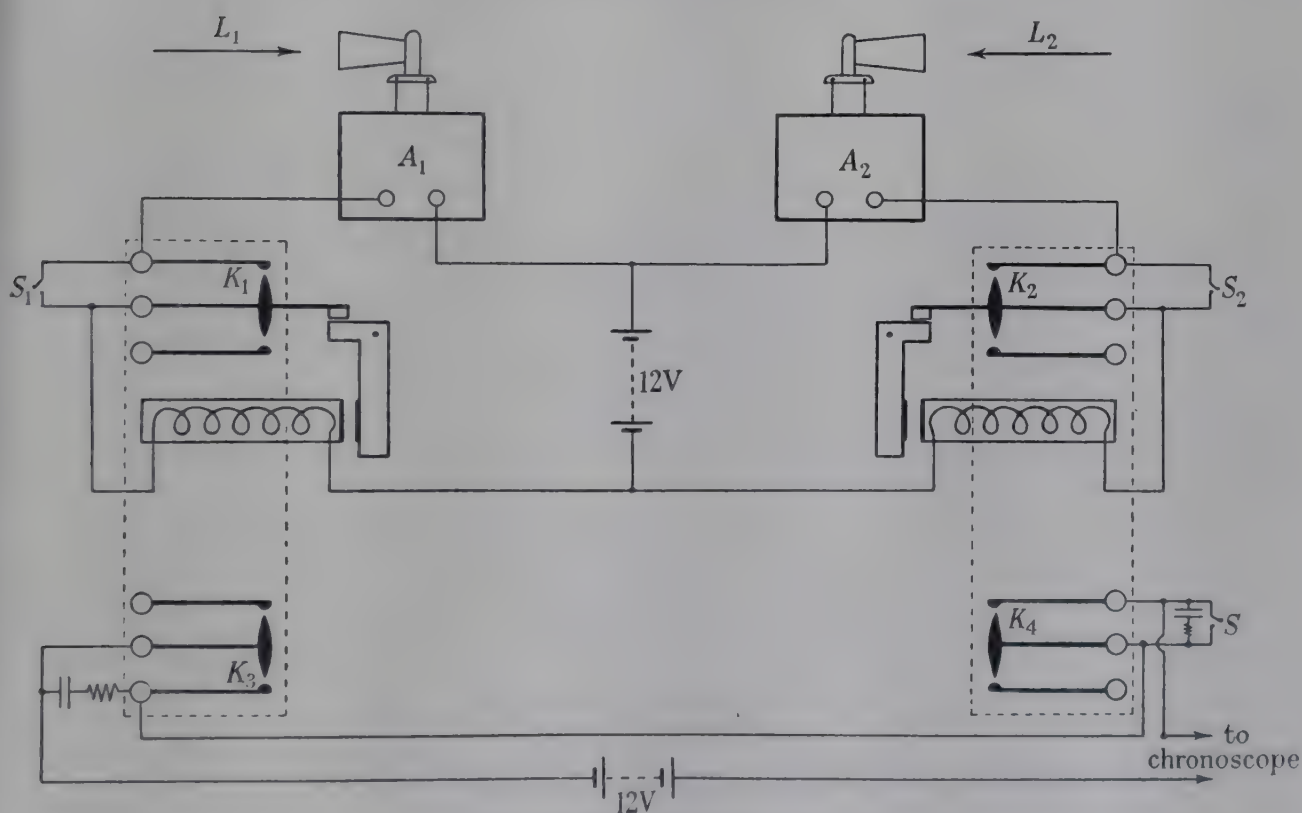


FIGURE 3. Circuit diagram.

Since the operation of the synchronous timer requires its circuit to be permanently made at the beginning of the time interval and to be permanently broken at the end of the interval, the *momentary* operation of the photocell relay must be translated into a permanent operation of a second relay. This is done by the use of locking relays as shown in the circuit diagram (figure 3), where it is seen that a momentary break at K_1 gives a permanent make at K_3 thus starting the timer, whilst a momentary

break at K_2 gives a permanent break at K_4 thus stopping the timer. S_1 and S_2 are set-reset switches, and S is a switch used to enable the second lamp source to be energized only a few seconds before the end of the time interval, an important consideration when the times of flow are about 20 min., since failure of the bulb due to continuous operation during a run would necessitate disturbance of the optical system. The times of operation of the mechanical relays employed are a few milliseconds, but since these times are reproducible to much less than a millisecond the 'personal error' of the system, even if as large as 0.01 sec., will always be the same to within a millisecond. The difference between two measured time intervals can thus be obtained correct to within about a millisecond.

MEASUREMENTS WITH A NON-POLAR LIQUID

As a typical non-polar liquid with well-known properties benzene was chosen, the specimen used being pure and dry. To begin with, runs were taken every half-hour and a systematic decrease in the time of flow was found in successive runs. The results of a typical series are shown in figure 4, where the circles and crosses represent the

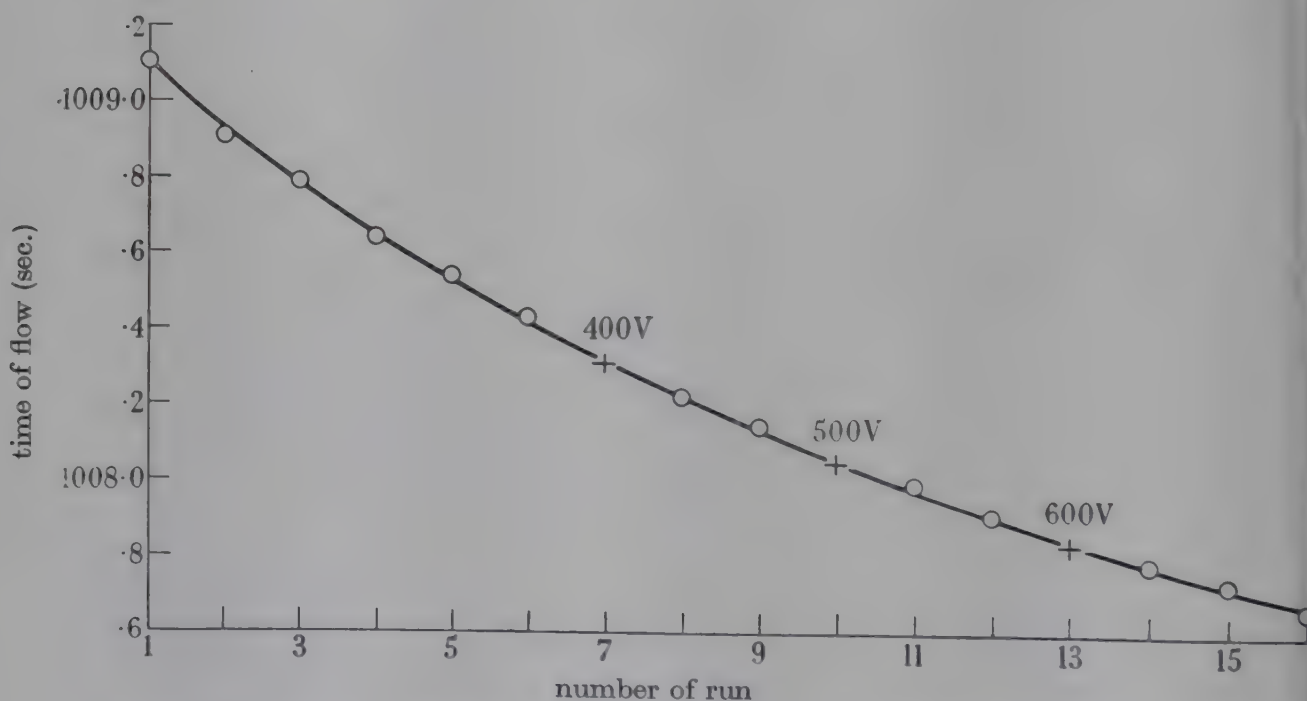


FIGURE 4. Benzene. Time of flow in successive runs. + indicates time of flow with 400, 500 and 600 V at 15,000 c./sec.

times of flow in sixteen successive trials. This decrease of viscosity, which at first appears surprising, is readily explained as due to the slight rise in temperature caused by the heat resulting from the work done in forcing the liquid through the narrow channel, both during the flow and during the resetting.

To show that the effect is of the magnitude required by this explanation, we note that, since the resetting is done under a pressure difference much greater than that prevailing during the measured flow, we can consider for rough purposes the heat

effects due to resetting alone. If the resetting is done under a head of h cm. of mercury, then the rise in temperature, supposing no heat is conducted away, is

$$d\theta = \frac{h \times 13.6 \times 981}{4.18 \times 10^7 \rho s},$$

where ρ is the density and s the specific heat of the liquid. The value of h being about 8 cm., while $\rho = 0.88$ g./cm.³, $s = 0.34$ for benzene, we have $d\theta = 0.0087^\circ$ C. The change in the time of flow is given by

$$\Delta t = t \frac{d\eta}{d\theta} \frac{d\theta}{\eta} = 0.14 \text{ sec.}$$

for benzene at 18° C, the time of flow t being about 1000 sec. The observed difference between the times of flow for the first two runs is about 0.16 sec., which is of the right magnitude. The systematic decrease in the difference of time of flow between successive runs is due to the fact that as the temperature rises the conduction rates of loss increase.

That the decreased time of flow is due to work-heating is confirmed by the fact that if, after a series of runs, the apparatus is left overnight, so that there is enough time for thermal equilibrium to establish itself, the time of flow returns to the value obtained in the first run of the previous day. Even with the runs at intervals of four hours, to which reference has already been made, the times of flow never differ by more than 0.01 sec., showing that the heat generated has been almost entirely conducted away.

The method adopted to measure the effect of an electric field is to take runs every half-hour and to apply the field for certain runs in the series. It is then possible by interpolation to estimate with precision what would have been the time of flow, without field, at the temperature prevailing during the flow with field for these particular runs and thus to see if any effect has been produced.

Thus, in the series for benzene represented in figure 4, 400 V at a frequency of 15,000 c./sec. was applied during the seventh run, 500 V during the tenth and 600 V during the thirteenth, the times of flow with field applied being represented by crosses. It will be seen that there is no change of time of flow exceeding $\frac{1}{100}$ sec. even for the highest field. Any change of viscosity of benzene due to an electric intensity of 42 kV/cm. does not therefore exceed 1 part in 100,000 or so. A series of runs taken with a frequency of 10,000 c./sec. led to a similar negative result.

The result of this work on benzene is not only to establish the negative effect to a very much higher degree of precision than hitherto but to show the consistent and generally satisfactory operation of the method.

POLAR LIQUIDS

(a) Monochlorbenzene

A pure *dry* specimen of monochlorbenzene, which is a polar liquid of extremely low conductivity and so does not show the steady voltage (spurious) effect referred to in the introduction (see also part I, p. 315) was subjected to an alternating field

of frequency 10,000 c./sec. The procedure was the same as in the case of benzene. A positive effect was obtained with p.d.'s of 100, 200 and 300 V, the results for 300 V, which are typical, being represented in figure 5.

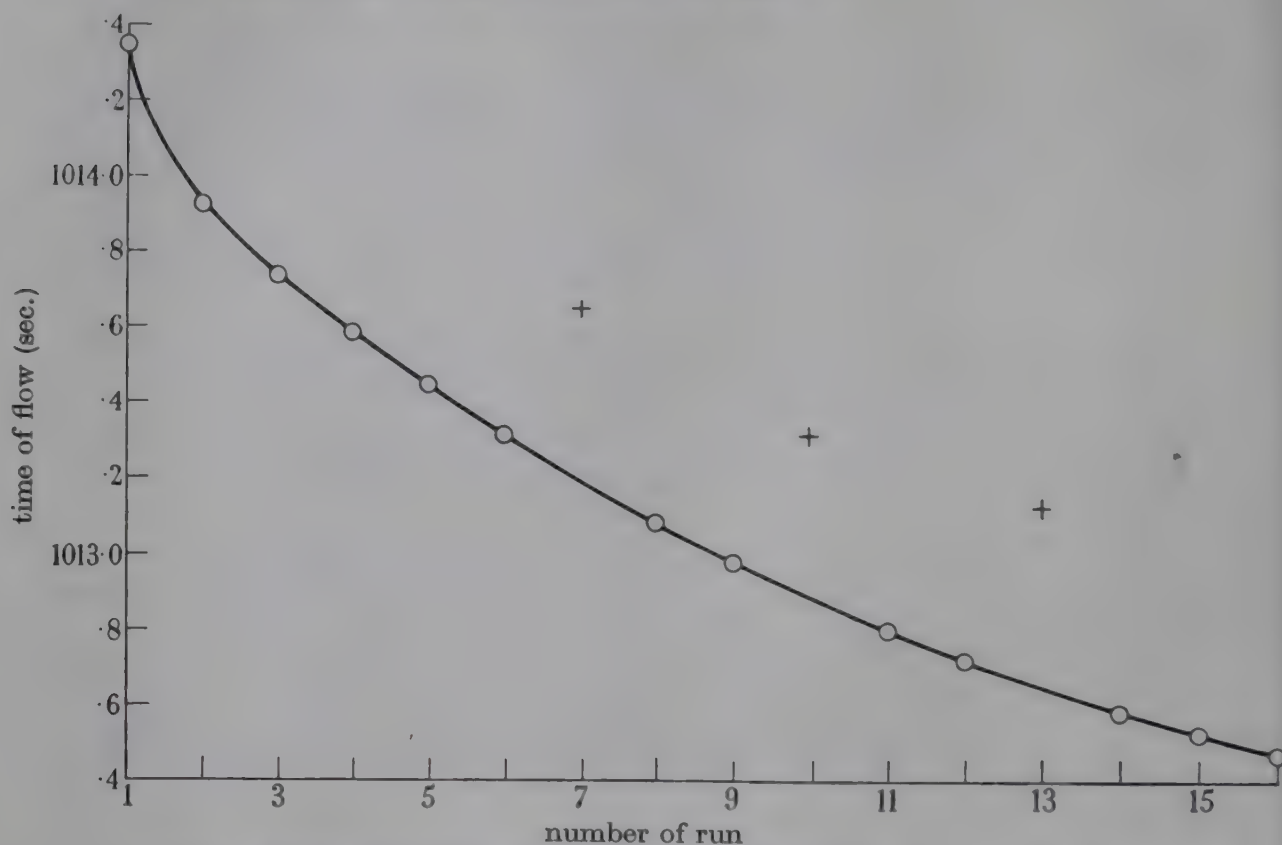


FIGURE 5. Monochlorobenzene. Time of flow in successive runs. + indicates time of flow with 300 V at 10,000 c./sec.

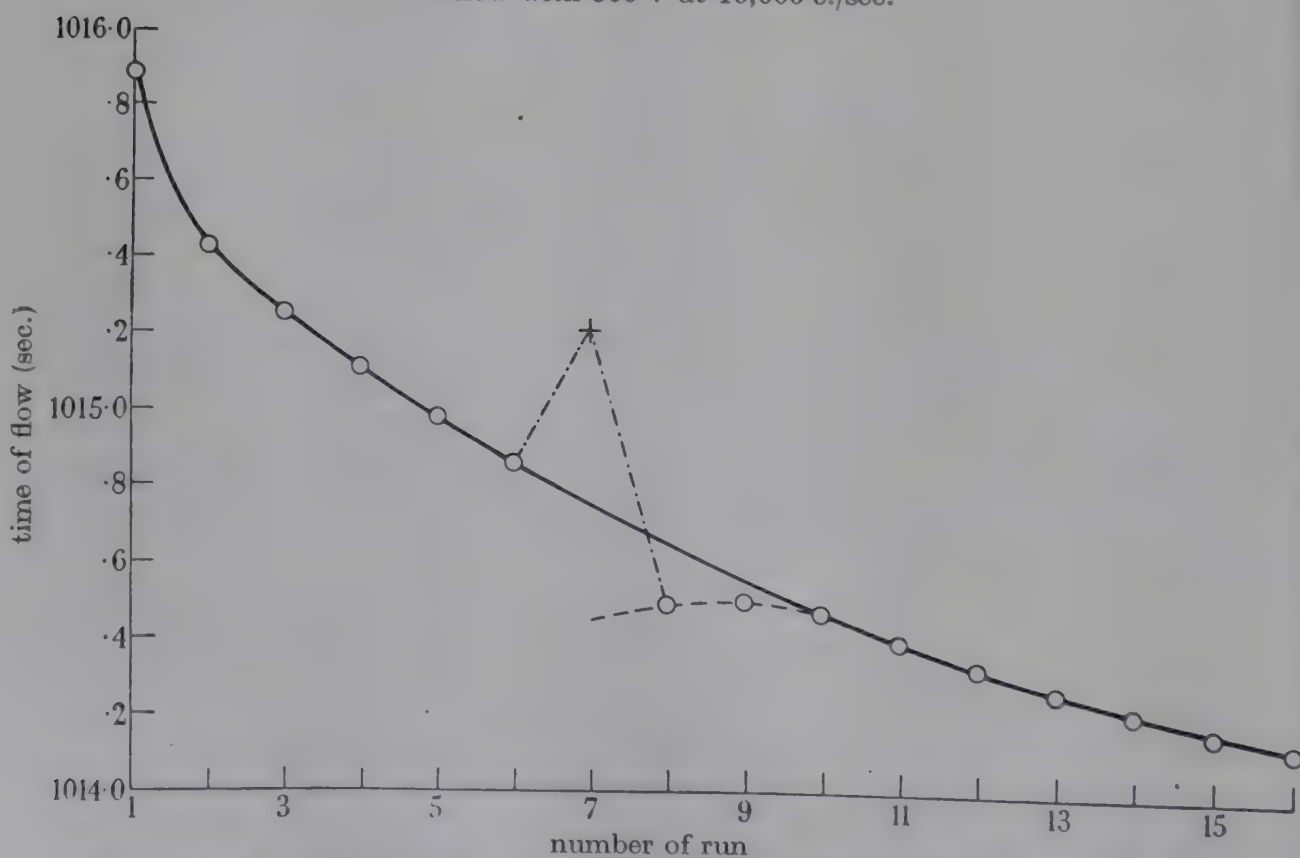


FIGURE 6. Monochlorobenzene. + indicates time of flow with 400 V at 10,000 c./sec.

The three applications of the field gave a consistent increase in time of flow of 0.197 sec. (average) on 1013 sec., the time of flow for the runs before and after the application of the field lying on a smooth curve showing the decrease discussed in the previous section. With higher voltages, however, the first few runs after the application of the field showed a systematic departure from the smooth curve, as shown in figures 6 and 7, which represent the observations with 400 and with 600 V.

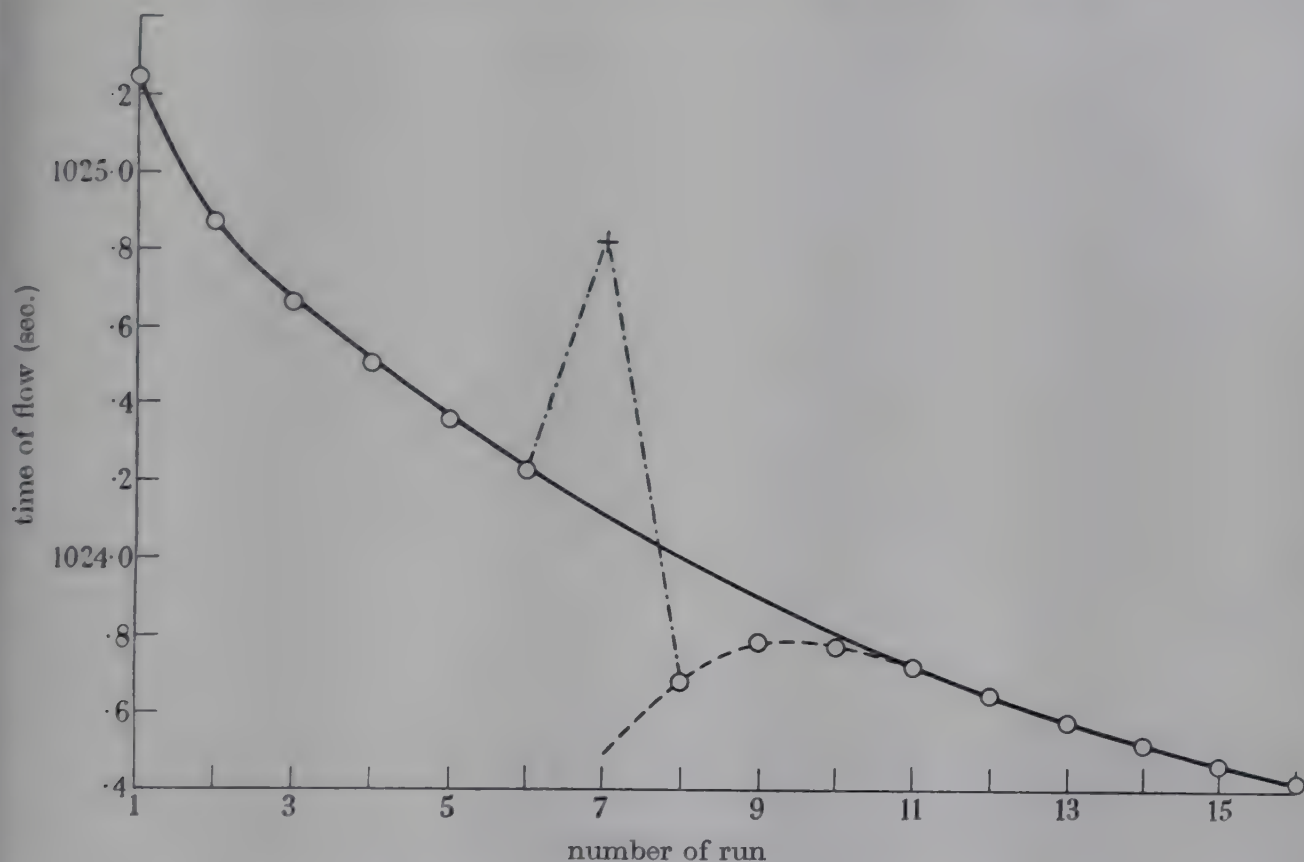


FIGURE 7. Monochlorbenzene. + 600 V at 10,000 c./sec.

The run just after the application of the field indicates a diminished viscosity which gradually recovers to normal during the next few runs, just as if the field had produced a slight heating of the liquid. Such dielectric heating is well known to occur in alternating fields (see Dodd (1948), where the particular case in question is discussed). Experiments carried out at the higher frequency of 15,000 c./sec. give a larger decrease of time of flow after the application of the field, as can be seen by comparing figure 8 with figure 7; this supports the view that the effect is due to dielectric heating. In figure 8, which shows the effect in its most exaggerated form, the heating is so large that the increase of viscosity due to the field still leaves the recorded time of flow below the normal curve.

At the higher voltages the effect was obtained by extrapolating the time of flow backwards, as shown by the dotted line, to find what it would have been, due to the dielectric heating, if there has been no direct effect of field on viscosity, and taking this as the zero from which to measure. When the extrapolation was not of the extreme character exemplified by the case of 600 V at 15,000 c./sec. the values found for the two frequencies agreed pretty well, as will be seen from table 1. The time

of flow, which has not been inserted, varied in different series of experiments from 1010 to 1035 sec., according to filling.

It is also clear from the table that the increase of viscosity is proportional to V^2 . The proportionality is well shown by figure 9. The constant of proportionality f in the equation $\frac{\Delta\eta}{\eta} = fV^2$ we propose to call the viscoelectric constant of the liquid.*

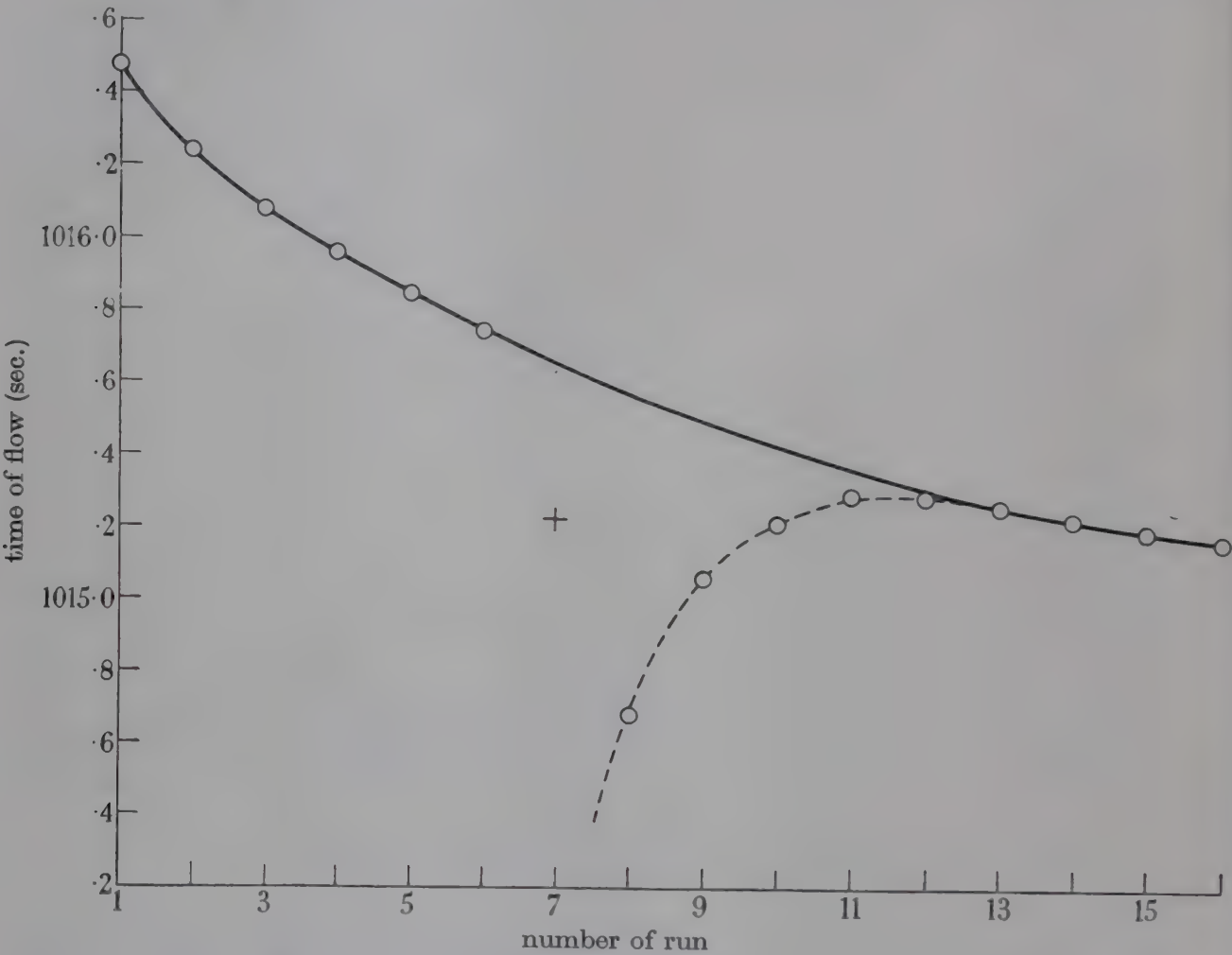


FIGURE 8. Monochlorbenzene. + 600 V at 15,000 c./sec.

TABLE 1. MONOCHLORBENZENE

peak voltage	r.m.s. voltage V	Δt		$10^4 \Delta t / t$	$10^3 \frac{\Delta t}{t} / V^2$
		10^4 c./sec.	1.5×10^4 c./set		
100	71	0.083	0.085	0.81	1.6
200	141	0.231	0.219	2.19	1.09
300	205	0.446	0.456	4.42	1.05
400	250	0.693	0.680	6.64	1.06
500	310	1.050	1.020	10.27	1.06
600	370	1.420	—	13.90	1.02

viscoelectric constant $f = 1.91 \times 10^{-7}$

* We originally proposed to call the constant the electroviscous constant. However, we find that the term 'electroviscous effect' has been applied to the change of viscosity of a suspension of small particles which results from the presence of an electric charge on the particles, an effect which has nothing to do with the phenomena investigated by us.

If F (the r.m.s. voltage V divided by the gap, 0.143 mm., across which the potential is applied) is expressed in electrostatic units, the value of f is 1.91×10^{-7} for monochlorobenzene.

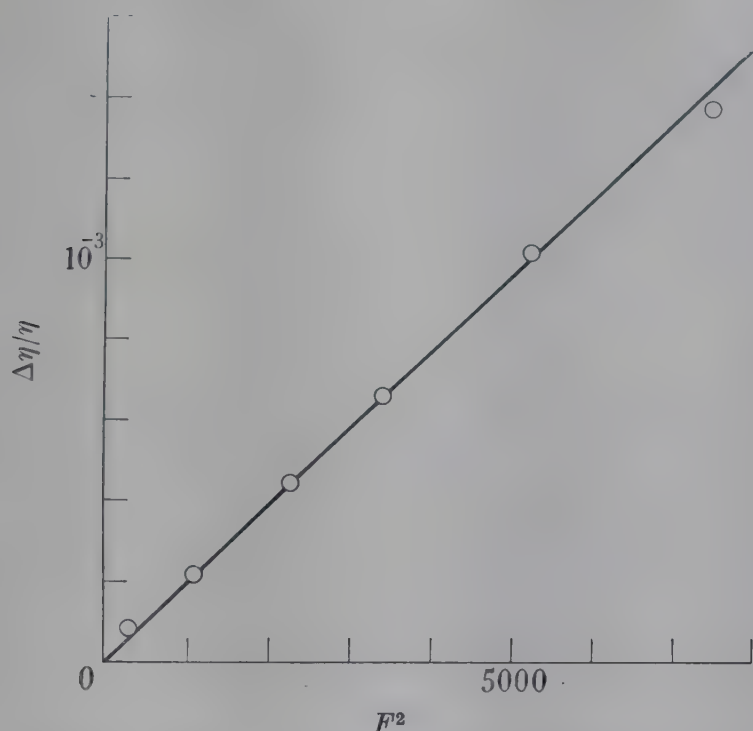


FIGURE 9. Monochlorobenzene. $\Delta\eta/\eta$ against F^2 (e.s.u.).

(b) Chloroform

A pure specimen of chloroform, which is a polar liquid that shows the steady voltage effect referred to in the introduction (see also part I, p. 316 and figure 14) was subjected to alternating voltages of frequency 10,000 c./sec. Heating effects due to the field, mainly dielectric heating but presumably also a little joulean heating, were detected for applied voltages in excess of 200 V and allowance was made for this

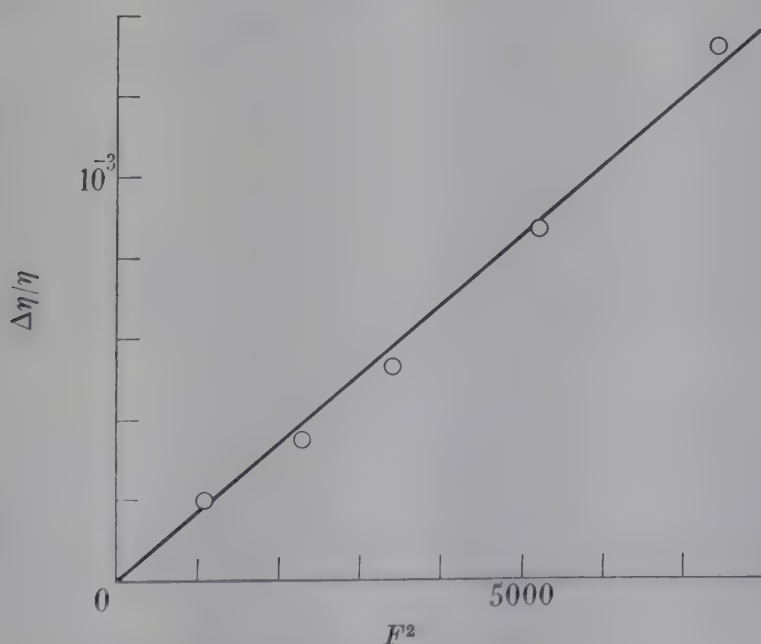


FIGURE 10. Chloroform. $\Delta\eta/\eta$ against F^2 (e.s.u.).

heating in the way described in the previous section. The results are shown in table 2 and figure 10.

TABLE 2. CHLOROFORM

peak voltage	r.m.s. voltage V	$10^4 \Delta t/t$	$10^8 \frac{\Delta t}{t} / V^2$
200	141	2.0	1.00
300	207	3.5	0.82
400	255	5.38	0.83
500	315	8.67	0.87
600	380	13.17	0.91

$f = 1.70 \times 10^{-7}$

(c) Amyl acetate

A pure specimen of amyl acetate, which is a polar liquid that shows the steady voltage effect (see part I, p. 313 and figure 9) was subjected to an alternating voltage of 10,000 c./sec. In this case the joulean heating is larger than with chloroform, the result being a rise of temperature, even at 300 V, which leads to a fall of viscosity when the field is applied, as shown in figure 11. Correction for this rise of temperature is made by extrapolation as indicated in the figures and previously described, and a rise of viscosity due to the field is thus obtained, but the consequence is that the measurements are less precise in this case, and cannot be made with peak voltages in excess of 500 V. The results are shown in table 3 and figure 12.

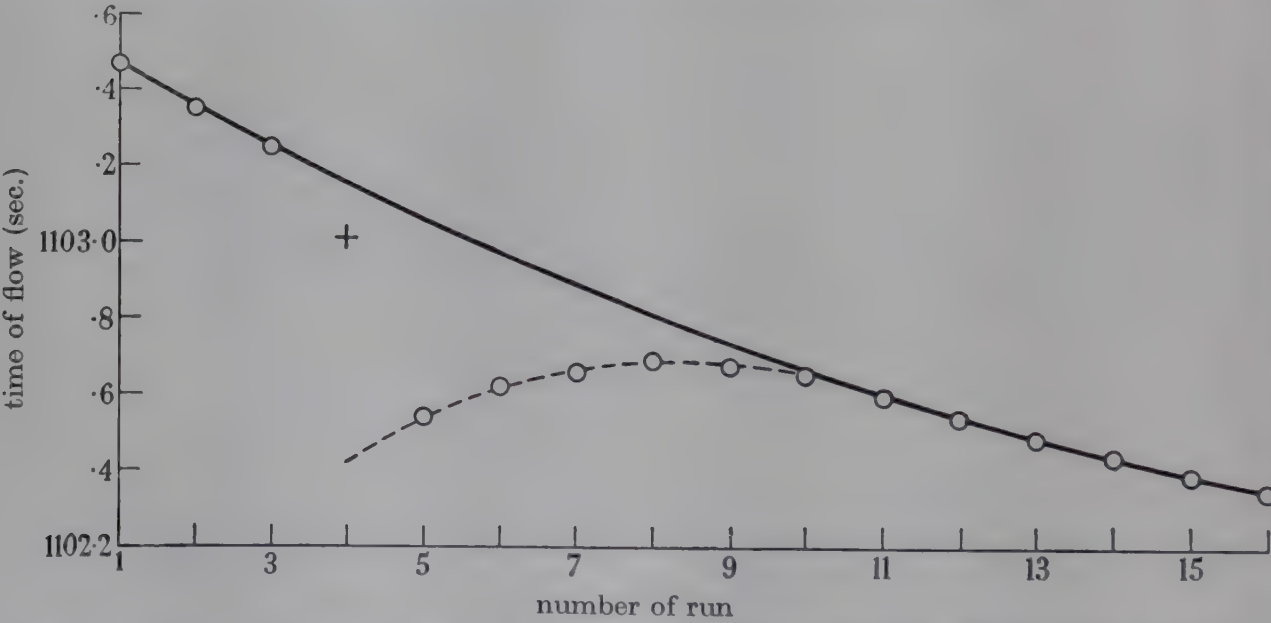
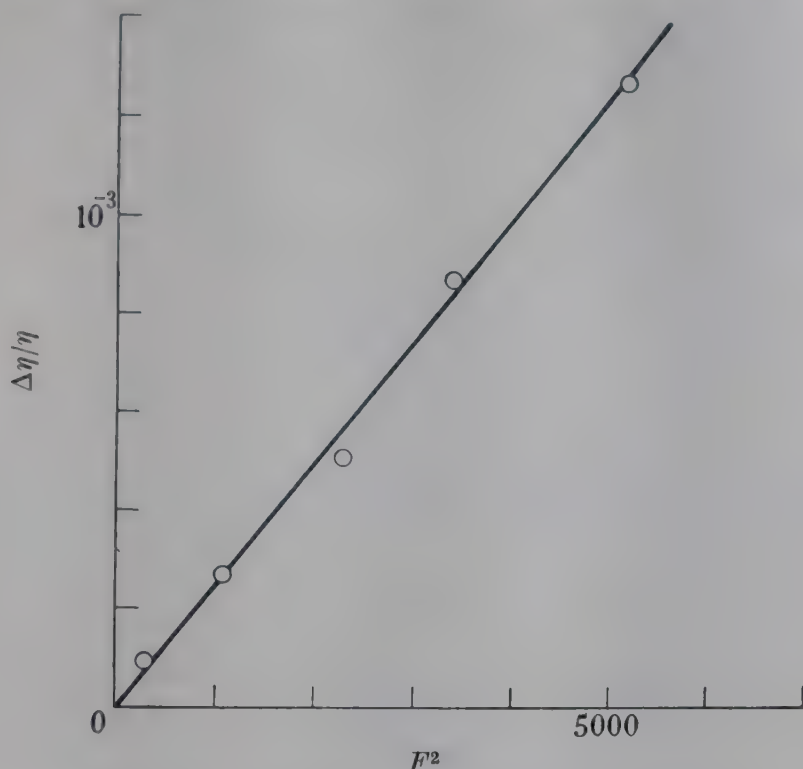


FIGURE 11. Amyl acetate. + 300 V at 10,000 c./sec., showing large heating effect of field.

TABLE 3. AMYL ACETATE

peak voltage	r.m.s. voltage V	$10^4 \Delta t/t$	$10^8 \frac{\Delta t}{t} / V^2$
200	141	2.7	1.3
300	205	5.06	1.20
400	250	8.72	1.39
500	310	12.74	1.32

$f = 2.46 \times 10^{-7}$

FIGURE 12. Amyl acetate. $\Delta\eta/\eta$ against F^2 (e.s.u.).

DISCUSSION

It is established, then, that for the three polar liquids tested there exists a true effect of electric field on the viscosity, but that no such effect exists for the typical non-polar liquid, benzene. Of the three polar liquids two show the large steady potential effect, due to secondary causes, described in part I, but the third, monochlorobenzene, shows this effect only when wet. The work of part I shows that no difference of behaviour between these liquids is to be expected under conditions suitable for establishing a significant effect, namely, a rapidly alternating voltage, and this anticipation was confirmed. The change of viscosity is proportional to the square of the intensity. The effect is a small one, only measurable with very sensitive methods, amounting to less than 1 part in 500 with the strongest fields.

When an electric field is applied to a liquid a small diminution of volume is to be anticipated, the effect being known as electrostriction, and at first it might seem possible that this change of volume might increase the viscosity in the same way as does increase of pressure, with corresponding compression. Ordinary thermodynamic considerations lead to the formula

$$\Delta v = -\frac{vF^2}{8\pi} \left(\frac{\partial \epsilon}{\partial p} \right)_{F, T}, \quad (1)$$

where v is the volume, F the intensity of the electric field (difference of potential between plates/distance between plates), ϵ the dielectric constant and p the pressure. Turning to the pressure effect, let the compressibility of the liquid be

$$\alpha = -\frac{1}{v} \frac{\partial v}{\partial p},$$

and the coefficient of change of viscosity with pressure be

$$\beta = \frac{1}{\eta} \frac{\partial \eta}{\partial p}.$$

Then

$$\left. \begin{aligned} \Delta p &= -\frac{1}{v\alpha} \Delta v, \\ \frac{\Delta \eta}{\eta} &= \beta \Delta p = -\frac{\beta}{v\alpha} \Delta v \\ &= \frac{1}{8\pi} \frac{\beta}{\alpha} \left(\frac{\partial \epsilon}{\partial p} \right)_{F,T} F^2, \end{aligned} \right\} \quad (2)$$

on the supposition that it is change of volume that produces the viscosity change.

Now, in the first place, α and β differ little between a typical non-polar liquid, such as benzene, and a polar liquid, such as chloroform (adiabatic compressibility = 6.5×10^{-11} cm. g.⁻¹ sec.² for benzene, 6.7×10^{-11} for chloroform; $\beta = 11.4 \times 10^{-10}$ cm. g.⁻¹ sec.² for benzene, 5.7×10^{-10} for chloroform*). The visco-electric effect is grossly different for the two classes of liquids. Secondly, the change of viscosity calculated from the formula (2) is of the wrong magnitude: for chloroform, for instance,

$$\frac{1}{8\pi} \frac{\beta}{\alpha} \left(\frac{\partial \epsilon}{\partial p} \right)_{F,T} = \frac{1}{8\pi} \frac{5.7 \times 10^{-10}}{6.7 \times 10^{-11}} 4.54 \times 10^{-10} = 1.54 \times 10^{-10}.$$

The experimental figure is 1.70×10^{-7} . The small change in volume due to the field cannot, then, be expected to have any direct bearing on the observed effect, and will be neglected.

Since the mechanism of the viscosity of liquids cannot yet be said to be satisfactorily explained, a detailed explanation of the effect established by us cannot be expected, but certain general observations seem possible. The effect of thermal agitation is to prevent a directional ordering of the polar molecules under the local field, such an ordering being, on the views put forward by one of us some time ago (Andrade 1934), favourable to the communication of momentum from molecule to molecule and so causing an increase in the viscosity. The field applied externally will tend to promote a directional ordering, which will be measured by the average moment in the direction of the field, which is $\frac{\mu^2}{3kT} F_i$,† the corresponding energy being $\frac{\mu^2}{3kT} F_i^2$. Here F_i is the field in the neighbourhood of the molecule caused by the externally applied field. If the electroviscous effect is γ times this quantity,

then

$$\gamma \frac{\mu^2}{3kT} F_i^2 = f F_i^2$$

and

$$f = \frac{\gamma \mu^2}{3kT},$$

* β is derived by extrapolation from Bridgman's figures.

† See, for example, Debye (1929).

where γ is of the dimensions $1/\text{energy} = 1/M$, say, M being of the order 10^{-16} erg. The effect may be considered as produced by a term added by the applied field to the energy factor in the formula

$$\eta = A e^{E/kT}.$$

Thus when the field is applied

$$\begin{aligned} \eta + \Delta\eta &= A \exp \left[\frac{E + \frac{1}{3}\gamma\mu^2 F_i^2}{kT} \right] \\ &= \eta \left(1 + \frac{1}{3} \frac{\gamma\mu^2 F_i^2}{kT} \right) \end{aligned}$$

if the added term is small.

The deviation of F_i from F has been the subject of much discussion, the treatment of the problem by Onsager (1936) being probably the most satisfactory to date. He derives the expressions

$$\bar{F}_i = \frac{3\epsilon}{2\epsilon + 1} F, \quad (3)$$

$$\mu_i = \frac{2\epsilon + 1}{2\epsilon + n^2} \frac{n^2 + 2}{3} \mu, \quad (4)$$

where μ_i is the effective moment in the condensed state, μ the moment measured for the dispersed state and n the refractive index. In table 4 the viscoelectric constant, recalculated for the internal field F_i , is given as f_i . Since $\frac{3\epsilon}{2\epsilon + 1}$ varies very little for values of ϵ from 4 to 6, f_i bears a nearly constant ratio to f for the liquids in question. The values of μ_i are calculated from (4), the value of n taken being that for sodium light. In the last column is given the value of M , the energy of magnitude 10^{-16} erg to which reference has already been made.

It is felt that, before attempting to take the theoretical discussion any further, systematic measurements of the viscoelectric constant for a variety of carefully chosen liquids are necessary. Such measurements will be made at the Royal Institution in the near future.

TABLE 4. TEMPERATURE 20° C

liquid	ϵ	$10^7 f$	$10^7 f_i$	$10^{18} \mu$	$10^{18} \mu_i$	$10^{23} \frac{\mu_i^2}{3kT}$	$10^{16} \frac{\mu_i^2}{3kT} / f_i$
chlorobenzene	5.94	1.91	1.00	1.69	2.21	4.03	4.03
chloroform	5.05	1.70	0.91	1.15	1.43	1.69	1.86
amyl acetate	4.88	2.46	1.33	1.82	2.20	3.99	3.00

APPENDIX

These experimental results allow us to make a very rough estimate of the change in the potential barrier concerned in flow which would be produced by a complete ordering of the polar molecules. It is obviously impossible to obtain anything like this complete ordering by the application of an external field, since electrical breakdown would ensue long before the field had reached sufficient magnitude, but the theoretical result is curious.

In (3) let us write the term due to flow as $b(F_i/F_0)^2$, where F_0 is a field strength whose choice will determine b . If we choose $F_0 = 3kT/\mu$, we have a field that corresponds to something like the complete ordering of the dipoles.*

E may be regarded as the potential barrier for the molecules oriented at random, while $1+b$ will be the barrier for molecules nearly completely ordered. To get the order of magnitude of b , we note that

$$b = \frac{\Delta\eta}{\eta F_i^2} F_0^2 \frac{kT}{E}.$$

For chloroform

$$\frac{\mu^2}{3kT} = 1.0 \times 10^{23},$$

while

$$\frac{\Delta\eta}{\eta} \frac{1}{F^2} = 1.7 \times 10^{-7},$$

$$F_0 = 1.1 \times 10^5$$

and

$$\frac{E}{kT} = 3$$

all values being approximate. Hence

$$b = \frac{1.7 \times 10^{-7}}{3} \times 1.2 \times 10^{10} = 680,$$

so that, making every allowance for the roughness of the approximations, the height of the barrier would be multiplied by several hundreds could complete ordering be achieved. The meaning, no doubt, is that solidification would take place long before that could occur.

We are indebted to Dr R. Eisenschitz for helpful discussion.

REFERENCES

- Andrade, E. N. da C. 1934 *Phil. Mag.* **17**, 497, 698.
 Andrade, E. N. da C. & Dodd, C. 1946 *Proc. Roy. Soc. A*, **187**, 296.
 Debye, P. 1929 *Polar molecules*. New York: Chemical Catalog Co.
 Dodd, C. 1948 *Proc. Phys. Soc.* **61**, 97.
 Onsager, L. 1936 *J. Amer. Chem. Soc.* **58**, 1486. (See also J. N. Wilson, 1939, *Chem. Rev.* **25**, 377.)

* At $\mu F/kT = 3$ Langevin's function is 0.67 of its limiting value.

Kinetics of the hydrogen/oxygen reaction

I. The explosion region in boric acid-coated vessels

BY SIR ALFRED EGERTON, F.R.S., AND D. R. WARREN

*Department of Chemical Engineering and Applied Chemistry, Imperial
College of Science and Technology, London*

(Received 14 August 1950)

The region of low-pressure explosion for hydrogen/oxygen mixtures in boric acid-coated vessels has been observed over a wide range of temperature, vessel size and mixture ratio. The results obtained cannot be explained by any of the existing mechanisms, and appear to conflict with the present interpretation of the effects of boric-acid coating.

The explosion boundary is described by a simple mathematical expression which demands the participation of a second-order chain-branching process not operative in salt-coated vessels. The previously excluded reaction



is postulated. A mechanism is outlined which successfully accounts for the shape and temperature dependence of the newly observed explosion region.

INTRODUCTION

The hydrogen/oxygen reaction has been very fully studied, and most workers are now in agreement upon at least the basic reactions of the mechanism (Lewis & von Elbe 1942; Hinshelwood 1946; Williams & Singer 1948).

The explosion limits have usually been considered in terms of pressure of auto-ignition as a function of temperature for a given mixture ratio, in which case the limits take the familiar form of an 'explosion peninsula'. If, however, the limits are plotted in terms of the partial pressures $[\text{H}_2]$ and $[\text{O}_2]$ at the explosion boundary, much can be learnt of the underlying reactions. The region then takes the form shown in figure 1A, the second limit appearing as a straight line deflected at the ends. The equation of the straight line is

$$[\text{H}_2] + a[\text{O}_2] = K, \quad (1)$$

which by the chain theory must signify an equality in the rates of chain branching and breaking. If the constant K represents the branching rate, then $[\text{H}_2] + a[\text{O}_2]$ is interpreted to mean chain destruction by triple collisions in the gas phase, an hypothesis confirmed by the fact that the value of a is approximately the calculated ratio of the collision numbers of the two gases (Grant & Hinshelwood 1933; Lewis & von Elbe 1942). The deflexions at extreme mixtures, which continue down to the first limit, are due to chain breaking at the walls. This process becomes the all-important factor at the first limit. An expression which describes the whole region has been given by Frost & Alyea (1933).

It was decided to examine the shape of the explosion region using various coatings on the vessel walls, so that from the degree of curvature of the second limit (or the value of the first limit if such could be reliably obtained) some information might be gathered as to the chain-destroying efficiencies of different surfaces.

Most salts tested gave an explosion region of the previously recorded type as in figure 1*A*, that is, the explosion limits approximated to the equation of Frost & Alyea if suitable constants were chosen. Vessels coated with boric acid and certain other materials, however, gave an entirely different curve (figure 1*B*) which could not be accounted for by any of the existing mechanisms, and which, furthermore, seemed to conflict directly with the present interpretation of limits in boric acid-coated vessels (Lewis & von Elbe 1942). Some hitherto unconsidered reaction was apparently in operation, and the kinetics of the reaction in boric acid-coated vessels were therefore deemed worthy of fuller investigation.

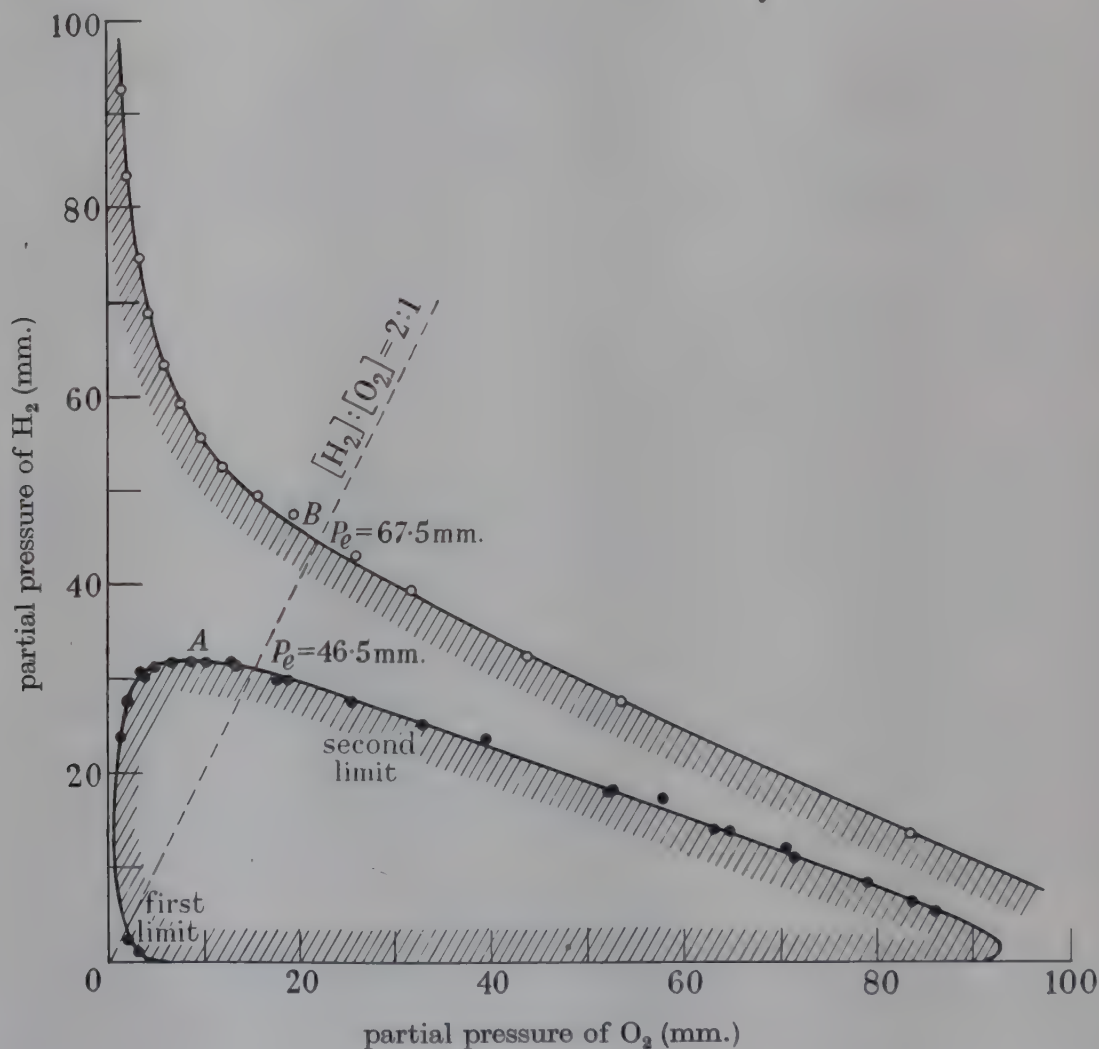


FIGURE 1. Explosion region of H_2/O_2 in a 7.4 cm. diameter vessel at 500°C . *A*, with KCl coating. *B*, with B_2O_3 coating. (P_e = total pressure at explosion limit.)

EXPERIMENTAL TECHNIQUE

A spherical Pyrex reaction vessel with capillary inlet was cleaned, then rinsed with saturated C.P. boric acid in 50 % alcohol and lightly drained. This was then warmed under vacuum to evaporate the solvent, leaving a translucent white deposit on the walls. At high temperatures this melted and remained transparent even on cooling again.

The vessel was maintained in an air-circulating furnace controlled by a gas-expansion thermostat; this has been fully described elsewhere (Warren 1951). The present work was carried out before the adoption of the sleeve mentioned in that paper, with the result that temperature gradients of some 3 to 4° C were detectable. Temperature was measured by a platinum to platinum-rhodium thermocouple in conjunction with a potentiometer and mirror galvanometer, and was maintained within $\pm 0.2^\circ\text{C}$ during any one run.

Cylinder hydrogen and oxygen were dried over calcium chloride and phosphorus pentoxide. To observe a limit, oxygen (or hydrogen) was let into the reaction vessel to a pressure well above that at which auto-explosion would occur. The other gas was then admitted (by a predetermining pipette method when speed was required) until the desired mixture ratio was obtained, and the evacuation begun immediately. The pressure at which explosion occurred was read directly on a mercury manometer. Between each run the vessel was evacuated to about 10^{-3} mm. Hg and swept out once with the first gas to be admitted in the next run.

For all but extreme mixture ratios, the rate of evacuation was found to have little effect upon the observed limits. Nevertheless, at high temperature it was carried out as quickly as was consistent with accurate manometer reading, to minimize any suppression by water formed in the slow reaction. Some difficulty, however, was encountered with very weak mixtures. If, in these cases, evacuation was allowed to occur quickly, it was found that the limits so obtained were low and approximately the same as in a KCl vessel, while too slow evacuation caused a sluggish ignition, rather difficult to observe exactly. It is for this reason that the last few points recorded for 540 and 550° C are of very doubtful accuracy, and the apparent 'turning-in' of the curves (shown in figure 2 as dotted lines) may quite possibly have no kinetic significance. Only the higher points observed have been recorded. Since at the very most this effect appears to be a secondary one, it will not be considered in the general treatment of the limit curve, but will be discussed separately.

It was found that if the vessel were filled with H_2 to a pressure, which, though high, was still within the explosion region (e.g. 80 mm. Hg at 500° C), then, if O_2 were quickly added to it, no explosion occurred, even though the addition caused a passage through the explosion region.

RESULTS

Temperature effects

Limits between 400 and 550° C are plotted in figure 2. Observations below 400° C were limited by feebleness of explosion and errors in manometer reading. Those above 550° C were prevented by the fact that in mixing the gases one strikes the third limit which tends to shatter the vessel. No first limit could be observed.

Vessel-size effects

A selection of the points observed in vessels of different sizes at 500 and 540° C is shown in figure 3. The difficulty found in observing limits with weak mixtures

becomes more noticeable in small vessels, the 'kick' barely registering on the manometer. At the same time one also meets increasing error due to 'dead space', less accuracy in mixing the gases, and the possibility of a slight temperature error, due to the gradients in the furnace. Nevertheless, the results obtained fail to show any significant dependence upon vessel size for these limits.

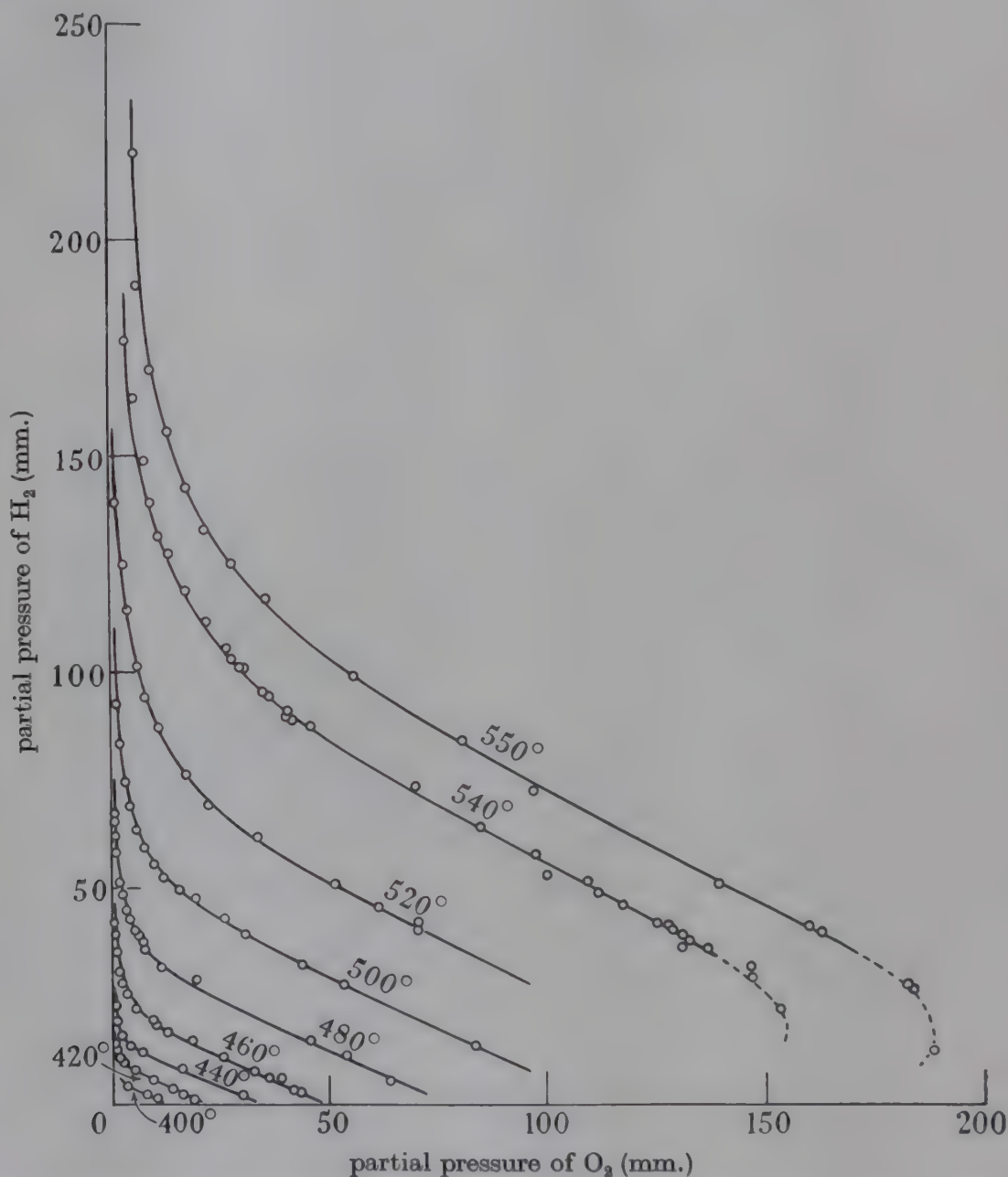


FIGURE 2. Explosion limit of H_2/O_2 in 7.4 cm. diameter Pyrex vessel coated with boric acid: 400 to 550° C.

Mathematical analysis

Each of the curves (neglecting the 'turning-in' for very weak mixtures as previously explained) is found to be very closely described by an expression of the type

$$[\text{H}_2] + a[\text{O}_2] = K + b[\text{O}_2]^{-\frac{1}{2}}. \quad (2)$$

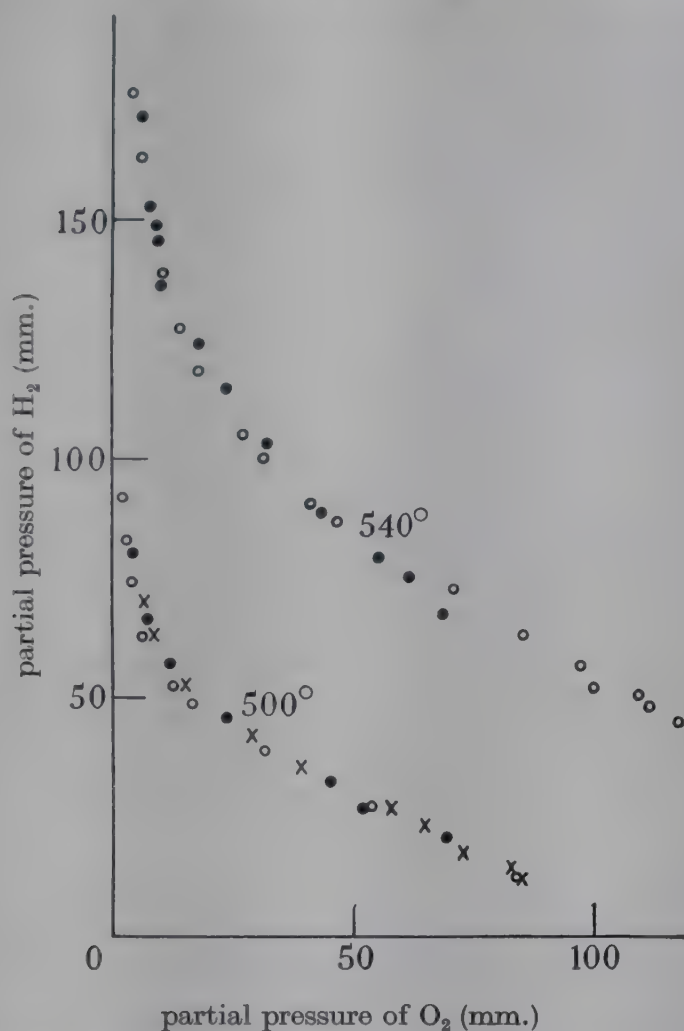


FIGURE 3. Variations in vessel size (B_2O_3 coating) at 500 and 540° C. $\circ$ 7.4 cm. diam. $\times$ 3.5 cm. diam. $\bullet$ 2.3 cm. diam.

The values of the constants for each temperature are obtained by selecting three well-spaced points on the curve and substituting. The results are as follows:

temperature (° C)	a	b	K
400	0.36	1.67	4.15
420	0.36	4.69	6.93
440	0.35	9.58	10.8
460	0.37	20.7	15.8
480	0.37	36.6	25.9
500	0.38	71.5	37.2
520	0.39	128.0	52.5
540	0.40	230.0	71.6
550	0.41	285.0	83.4

If we now plot $\log b$ and $\log K$ against $1/T$ as in figure 4, we find

$$b = 4.7 \times 10^{11} \times e^{-34,600/RT} \quad \text{and} \quad K = 6.3 \times 10^7 \times e^{-22,000/RT}.$$

The constant a merely increases slightly with rising temperature. Previous workers have observed values over the range 0.33 to 0.4 (Williams & Singer 1948).

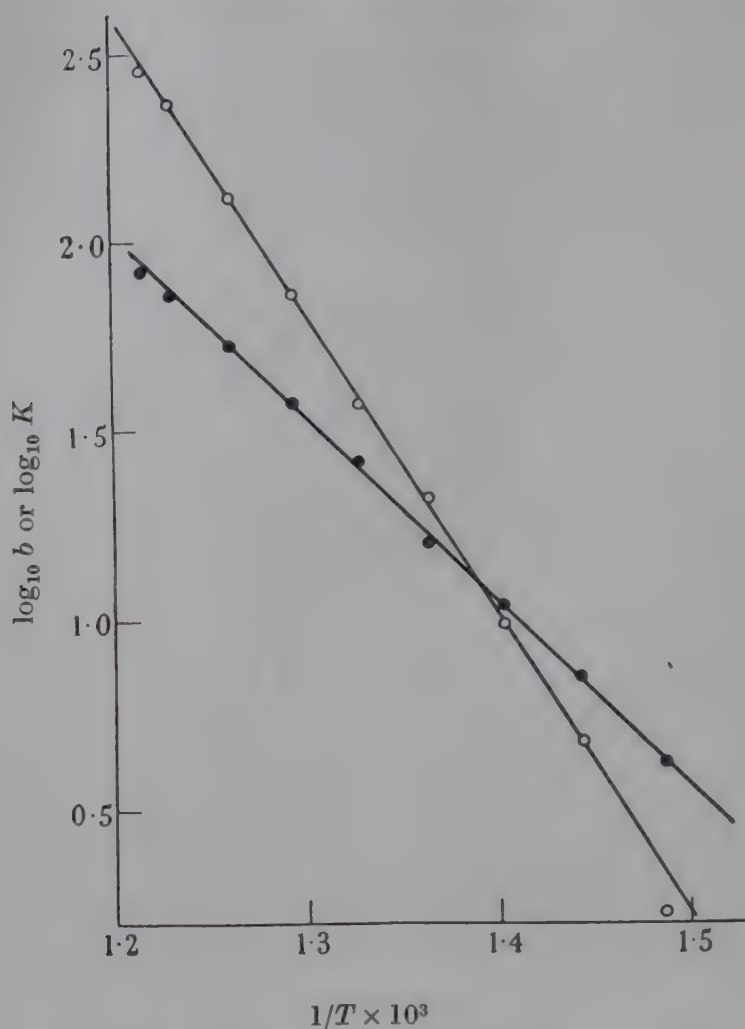


FIGURE 4. Variations of the constants b (shown ○) and K (shown ●) with temperature.

DISCUSSION

Comparison with earlier work

Explosions in B_2O_3 -coated vessels have previously been recorded by Lewis & von Elbe (1942), who find for a $2H_2 + O_2$ mixture at $530^\circ C$ a limit of 111 to 115 mm. With this we are in agreement, finding 112 mm. independent of vessel size. They have not recorded limits for other mixtures and have evidently assumed the limits to be straight lines as in KCl vessels. For the limits in small KCl vessels they give (quoting their symbols)

$$[M]_e = 2 \frac{k_6}{k_2} \text{ (their equation (20));}$$

for large KCl or $K_2B_4O_7$ vessels

$$[M]_e = 3 \frac{k_6}{k_2} \text{ (their equation (19) and p. 383),}$$

and for B_2O_3 vessels

$$[M]_e > 3 \frac{k_6}{k_2} \text{ (their p. 383).}$$

The term $[M]_e$ is by definition a measure of $[H_2] + a [O_2]$ at the explosion limit, where a is the relative third-body efficiency of O_2 as previously mentioned. Thus in the

co-ordinates used here these equations give limits which are parallel straight lines cutting off intercepts on the $[H_2]$ axis in the ratio 2:3: > 3.

It is agreed that this does happen to represent the approximate ratio of the second limits in KCl and B_2O_3 vessels for a 2:1 mixture such as Lewis & von Elbe used (cf. figure 1: $P_{eKCl}:P_{eB_2O_3} = 46.5:67.5 \simeq 2:3$). However, it does not do so for other mixture ratios, the B_2O_3 limits being far from straight lines.

Comparison with limits in KCl

If the term $b[O_2]^{-\frac{1}{2}}$ is omitted from equation (2), one is left with the straight line

$$[H_2] + a[O_2] = K.$$

This appears to be identical with the straight line for the upper limit in KCl vessels. For example, at 500° C (figure 1) the B_2O_3 limit is found to be given by

$$[H_2] + 0.38[O_2] = 37 + 71.5[O_2]^{-\frac{1}{2}},$$

while the KCl vessel upper limit is essentially

$$[H_2] + 0.37[O_2] = 38.$$

The two cases, then, apparently have the same rate of chain branching (K), and the same rate of chain breaking by triple collisions, as expressed by $[H_2] + a[O_2]$. In the KCl vessel one has to add another term (see Frost & Alyea, 1933, equation (5)) to account for the turning down of the limits due to chain breaking at the walls. This effect is absent in the observed B_2O_3 limits, and the new term $b[O_2]^{-\frac{1}{2}}$ must be added, which, since it is on the K side of the equation, must represent a new chain-branching process not operative in KCl vessels.

The form of this term $b[O_2]^{-\frac{1}{2}}$ is such that it could not be due to another first-order branching process, which would merely give a constant as in the case of K ; nor, being independent of vessel size, is it likely to be a chain production at the walls. A gas-phase reaction with B_2O_3 seems impossible due to the very small vapour pressure of this substance at these temperatures (Speiser 1950). On the other hand, the term is exactly as one would expect for the participation of a branching process of the second order with regard to chain carriers. For, in a system in which there are n chain particles, let

$$n_0 = \text{rate of initiation,}$$

$$pn = \text{rate of breaking (1st order),}$$

$$qn = \text{rate of branching (1st order),}$$

$$rn^2 = \text{rate of branching (2nd order).}$$

$$\text{Then } dn/dt = n_0 - pn + qn + rn^2.$$

$$\text{In the stationary state } dn/dt = 0$$

$$n = \frac{p - q \pm \sqrt{\{(p - q)^2 - 4rn_0\}}}{2r}. \quad (3)$$

If $4rn_0 > (p-q)^2$, then n is unreal, i.e. there is no stationary state and explosion results. Thus the boundary of the explosion region is defined by

$$4rn_0 = (p-q)^2, \quad p = q + (4rn_0)^{\frac{1}{2}}. \quad (4)$$

This is identical with equation (2), for $[H_2] + a[O_2]$ has already been identified with first-order chain breaking (p), and K with first-order branching (q).

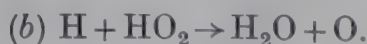
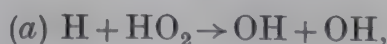
Necessary modification of the present mechanism

Lewis & von Elbe have given a very highly developed mechanism which successfully accounts for much of our knowledge of the H_2/O_2 reaction. An effort will therefore be made in accounting for these new facts to depart from that mechanism as little as possible. However, a number of the reactions (e.g. those included to account for wall effects and slow reaction) need not be considered at this stage.

Our main amendment to the present scheme is the addition of a second-order chain-branching process. Any collision between two of the active particles H, O or OH to produce three active particles is inconceivable. Therefore one is limited to consideration of reactions involving the inactive particle HO_2 , which does not for the most part reproduce chains. H_2O_2 cannot be considered as a chain particle, since, according to the Lewis & von Elbe mechanism and to the work of Pease (1930), it reaches a steady-state concentration which is proportional to $[H_2]$ and independent of chain carrier concentration.

The probability of a reaction involving HO_2 is confirmed by certain other considerations. B_2O_3 surfaces have been used for preserving peroxidic substances (cf. McLane 1949), while KCl appears to destroy them very efficiently (cf. Pease 1930); hence the second-order effect would not be noticeable in KCl vessels. Again, observations of the slow reaction above the second limit show that a considerable induction period occurs before chain carriers reach their steady-state concentrations, this delay being longest in weak mixtures. This would account for the fact that sudden manipulation with weak mixtures can give limits about the same as in KCl vessels, and also, as previously described, that a sudden passage through the upper part of the explosion region will not result in explosion.

Reactions of either O or OH with HO_2 to give two or more active particles seem improbable, and in any case appear to lead to an unsatisfactory result in the subsequent calculation. One is thus left to consider the reactions (both exothermic)



Since all O atoms disappear by the reaction



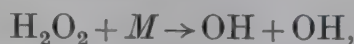
and all OH radicals by



then the two processes (a) and (b) are, for kinetic purposes, identical. On the grounds of being the more probable, the former will be chosen for future reference.

This particular reaction has received some consideration from Lewis & von Elbe (1942), and they have come to the conclusion that, together with all second-order processes, it can only contribute to a negligible extent. Most of their data has, however, been derived with KCl-coated vessels, in which such a reaction would be suppressed, while that which they derive from clean and B_2O_3 -coated vessels would appear to require reconsideration in the light of these results. The reaction has recently been postulated in accounting for the reaction of H atoms and molecular oxygen (Badin 1948).

One other modification must be made. By a comparison of equations (2) and (4), the constant b is seen to be also a function of the chain initiation process n_0 . This in the Lewis & von Elbe mechanism is given by



where H_2O_2 is maintained by the chain process at a stationary state proportional to the hydrogen concentration. (The alternative scheme of Hinshelwood also involves an M term.) However, b is found to be a simple Arrhenius function of temperature without any dependence upon $[M]$, even though the measurements have been made over a wide range of pressures (8 to 160 mm. Hg). We would therefore substitute the reaction



This modification merely demands that the lifetime of the energy-rich $H_2O_2^*$ molecule is long compared with the time between collisions at the pressures concerned. This seems quite reasonable, as the energy may be distributed among the three available bonds for some time before sufficient is localized in the central bond to cause dissociation.

Development of the mechanism

While it is admitted that other reactions may have to be added to account for other facts, such as the slow reaction or the third limit, the following will be considered as the simplest mechanism capable of satisfying the data recorded in this paper:

1. $H + O_2 \rightarrow OH + O.$
2. $O + H_2 \rightarrow OH + H.$
3. $OH + H_2 \rightarrow H_2O + H.$
4. $H + O_2 + M \rightarrow HO_2 + M.$
5. $HO_2 \xrightarrow{\text{wall}} \frac{1}{2}H_2O_2 + \frac{1}{2}O_2.$
6. $H_2O_2 + HO_2 \rightarrow H_2O + O_2 + OH.$
7. $H_2O_2 \rightarrow OH + OH.$
8. $H + HO_2 \rightarrow OH + OH.$

Since all O atoms are destroyed by reaction 2 only and all hydroxyls by reaction 3 only, computation can be simplified by condensing the reactions down to:

1. $H + O_2 \rightarrow \rightarrow 2H_2O + 3H.$
4. $H + O_2 + M \rightarrow HO_2 + M.$
5. $HO_2 \xrightarrow{\text{wall}} \frac{1}{2}H_2O_2 + \frac{1}{2}O_2.$
6. $H_2O_2 + HO_2 \rightarrow \rightarrow 2H_2O + O_2 + H.$
7. $H_2O_2 \rightarrow \rightarrow 2H_2O + 2H.$
8. $H + HO_2 \rightarrow \rightarrow 2H_2O + 2H.$

The rate of reaction 5 is $K_5[\text{HO}_2]$, where K_5 will be a function of vessel size, wall efficiency and the diffusion coefficient of the mixture, but, as this factor does not appear in the final result, no approximation need be made.

In this, as in previous work, M refers to any molecule capable of taking part in the triple collision and hence $[M] = [\text{H}_2] + a[\text{O}_2] + a'[\text{X}] + \dots$ where $a, a', \dots$ represent the relative third-body efficiencies of the respective molecules.

At the stationary state:

$$\frac{d\text{H}}{dt} = 2k_1[\text{H}][\text{O}_2] - k_4[\text{H}][\text{O}_2][M] + k_6[\text{H}_2\text{O}_2][\text{HO}_2] + 2k_7[\text{H}_2\text{O}_2] + k_8[\text{H}][\text{HO}_2] = 0, \quad (5)$$

$$\frac{d\text{HO}_2}{dt} = k_4[\text{H}][\text{O}_2][M] - K_5[\text{HO}_2] - k_6[\text{H}_2\text{O}_2][\text{HO}_2] - k_8[\text{H}][\text{HO}_2] = 0, \quad (6)$$

$$\frac{d\text{H}_2\text{O}_2}{dt} = \frac{1}{2}K_5[\text{HO}_2] - k_6[\text{H}_2\text{O}_2][\text{HO}_2] - k_7[\text{H}_2\text{O}_2] = 0. \quad (7)$$

From equation (7)
$$[\text{H}_2\text{O}_2] = \frac{K_5[\text{HO}_2]}{2(k_6[\text{HO}_2] + k_7)}. \quad (8)$$

Adding equations (5), (6) and twice (7)

$$[\text{H}] = \frac{k_6[\text{H}_2\text{O}_2][\text{HO}_2]}{k_1[\text{O}_2]}, \quad (9)$$

whence, from (8) and (9),

$$[\text{H}] = \frac{k_6 K_5 [\text{HO}_2]^2}{2k_1 [\text{O}_2] (k_6 [\text{HO}_2] + k_7)}. \quad (10)$$

*Substituting (8) and (10) into equation (6) and dividing by $K_5[\text{HO}_2]$ one obtains the following expression for $[\text{HO}_2]$:

$$\frac{k_8 k_6}{[\text{O}_2]} [\text{HO}_2]^2 - (k_4 k_6 [M] - 3k_1 k_6) [\text{HO}_2] + 2k_1 k_7 = 0. \quad (11)$$

Writing this as
$$l[\text{HO}_2]^2 - m[\text{HO}_2] + n = 0,$$

one obtains
$$[\text{HO}_2] = \frac{m \pm \sqrt{(m^2 - 4ln)}}{2l}.$$

The HO_2 concentration, then, will be built up by the chain process to a real steady-state value if $m^2 > 4ln$, and a slow reaction will occur. If, however, $m^2 < 4ln$, this steady-state concentration is unreal, and explosion will take place. The explosion limit is therefore given by

$$m = (4ln)^{\frac{1}{2}},$$

i.e.
$$k_4 k_6 [M] - 3k_1 k_6 = \left(\frac{8k_8 k_6 k_1 k_7}{[\text{O}_2]} \right)^{\frac{1}{2}}$$

or
$$[M] = [\text{H}_2] + a[\text{O}_2] = \frac{3k_1}{k_4} + \left(\frac{8k_8 k_1 k_7}{k_6 k_4^2} \right)^{\frac{1}{2}} [\text{O}_2]^{-\frac{1}{2}}. \quad (12)$$

Comparison of theory with data

Equation (12) is identical with the experimentally determined equation (2) if

$$\frac{3k_1}{k_4} = K (= 6.3 \times 10^7 \times e^{-22,000/RT})$$

and

$$\left(\frac{8k_8k_1k_7}{k_6k_4^2}\right)^{\frac{1}{2}} = b (= 4.7 \times 10^{11} \times e^{-34,600/RT}).$$

For the activation energies of reactions 1 and 4, Frost & Alyea have found (using KCl vessels) $E_1 = 24,200$ cal., and $E_4 =$ approx. 2000 cal. The value $E_1 - E_4 = 22,000$ cal. given above is therefore in good agreement with their results.

A rough check can be made on this interpretation of the constant b if one is willing to make estimates of the activation energies of reactions 1, 4, 6, 7 and 8.

For reactions 1 and 4 we will accept the values described above.

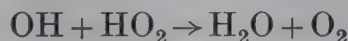
For reaction 7 the dissociation energy of H_2O_2 is taken to be 52,000 cal. (Giguère 1947).

Reactions 6 and 8 are exothermic free-radical reactions and their activation energies will be small and probably of the same order, so that, as a first approximation, they may be considered to cancel. Thus

$$\begin{aligned} b &\propto \exp - [(24,200 + 52,000 - 2000 - 2000)/RT]^{\frac{1}{2}} \\ &\propto \exp [-36,100/RT]. \end{aligned}$$

Considering the approximations made, this is in reasonable agreement with the observed value of 34,600, especially as reaction 6 will almost certainly have a higher activation energy than the bi-free-radical reaction 8.

A consideration might be made of any possible explanation the mechanism affords for the 'turning in' of the limits if such should prove significant. The effect is evidently due to some chain-destroying process inversely proportional to $[H_2]$ and not operative in KCl vessels. As $[OH]$ is inversely proportional to $[H_2]$, the reaction



is suggested as a possibility.

With regard to the present controversy over chain initiation (see Lewis & von Elbe 1948), one should perhaps point out that this mechanism provides general support for the conception of chain initiation as a self-sustained process, rather than being independent of the mechanism as in the hydrogen dissociation theory. It also provides indirect evidence that the dissociation of hydrogen peroxide is a monomolecular process down to pressures of a few millimetres.

One of us (D.R.W.) wishes to acknowledge his indebtedness to the Australian Department of Supply for granting the facilities by which this work was made possible.

REFERENCES

- Badin, E. J. 1948 *Third symposium on combustion, flame and explosion phenomena, Wisconsin*, p. 382. Baltimore: Williams and Wilkins.
 Frost, A. A. & Alyea, H. N. 1933 *J. Amer. Chem. Soc.* **55**, 3227.
 Giguère, P. A. 1947 *Canad. J. Res.* **25B**, 135.
 Grant, G. H. & Hinshelwood, C. N. 1933 *Proc. Roy. Soc. A*, **141**, 29.
 Hinshelwood, C. N. 1946 *Proc. Roy. Soc. A*, **188**, 1.

Lewis, B. & von Elbe, G. 1942 *J. Chem. Phys.* **10**, 366.

Lewis, B. & von Elbe, G. 1948 *Third symposium on combustion, flame and explosion phenomena. Wisconsin*, p. 484. Baltimore: Williams and Wilkins.

McLane, C. K. 1949 *J. Chem. Phys.* **17**, 379.

Pease, R. N. 1930 *J. Amer. Chem. Soc.* **52**, 5106.

Speiser, R. 1950 *J. Amer. Chem. Soc.* **72**, 2578.

Warren, D. R. 1951 *J. Soc. Chem. Ind., Lond.*, (in the Press).

Williams, G. & Singer, K. 1948 *Ann. Rep. Chem. Soc.* **45**, 51.

Binding energy of the triton

BY TSI-MING HU AND KUNG-NGOU HSU

University College, London

(Communicated by H. S. W. Massey, F.R.S.—Received 22 May 1950)

The variation method is employed to calculate the binding energy of the triton assuming charge-independent, two-body, Yukawa shape interactions between nucleons in which tensor forces are included. More complete trial wave functions are used than employed hitherto in such calculations, and it is found that an interaction of Yukawa shape with constants adjusted to fit the observed data on the binding energy, quadrupole moment and magnetic moment of the deuteron, the low-energy and high-energy scattering of neutrons by protons, the photodisintegration of the deuteron and the coherent scattering of slow neutrons gives an approximately correct binding energy for the triton. Calculations are also carried out with interactions of the same type but with different constants. The exchange character of the forces remains unimportant. It is confirmed that the difference in the binding energies of ^3H and ^3He can be ascribed to the effect of Coulomb repulsion between the protons in the latter nucleus. The wave functions found are used to compute the magnetic moments of the two nuclei but do not contain sufficient admixture of P component to explain the observed values.

1. INTRODUCTION

Until quite recently attempts to determine the nature of the fundamental interactions between nucleons from a largely empirical approach had met with very discouraging results. The number of adjustable constants which had to be included in the assumed interactions had to be increased to fit each new set of observed data. During the last year revised measurements of certain of the fundamental magnitudes determining the interaction have somewhat improved the prospect. It now seems that the interaction of the form proposed by Rarita & Schwinger (1941*b*), which involves four adjustable constants, can fit most of the observed data for low-energy interactions. This interaction is of the form

$$\mathcal{V}(r) = -\frac{\hbar^2}{M} \frac{a}{r_0^2} P \left\{ 1 + \frac{1}{2} g (\boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2 - 1) + \gamma \left(\frac{3 \boldsymbol{\sigma}_1 \cdot \mathbf{r} \boldsymbol{\sigma}_2 \cdot \mathbf{r}}{r^2} - \boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2 \right) \right\} J\left(\frac{r}{r_0}\right). \quad (1)$$

σ_1 and σ_2 are the spin operators for the two nucleons, $\mathbf{r}$ their vector separation and M their mass. a , r_0 , g and γ are adjustable constants and J is a shape function which determines the radial variation of the interaction. P is an exchange operator. For low-energy interactions the presence or otherwise of P is unimportant. For triplet

states γ is the ratio of the central to non-central component of the interaction, while in singlet states the non-central component vanishes leaving a central interaction which is $(1 - 2g)$ times as great as the central component for the triplet state.

It appeared from the analysis made by Blatt & Jackson (1949), Jackson & Blatt (1950) and by Bethe (1949) that, if the function J is taken to be of the Yukawa form

$$J(x) = x^{-1}e^{-x},$$

the interaction (1) is consistent with the most recent measurements of the following available at that time:

- (a) the binding energy of the deuteron,
- (b) the low-velocity limit of the neutron-proton scattering cross-section,
- (c) the coherent scattering of neutrons by crystals,
- (d) the quadrupole moment of the deuteron,
- (e) the magnetic moment of the deuteron,
- (f) the scattering of protons, with energies up to 10 MeV, by protons.

More recent measurements of the coherent scattering of slow neutrons (Hughes, Burgy & Ringo 1950) and of the cross-section for the photodisintegration of the deuteron (Barnes, Stafford & Wilkinson 1950) have partially destroyed the consistency achieved with the Yukawa shape, but with the exponential shape

$$J(x) = e^{-2x}$$

consistency is restored. Even with the Yukawa shape the present inconsistency is not very marked and may disappear in the light of further measurements. Meanwhile, the best choice of the four parameters for the Yukawa shape is found to be (Castillejo 1951)

$$a = 1.14 \times 10^{-13} \text{ cm.}, \quad r_0 = 1.36 \times 10^{-13} \text{ cm.}, \quad g = -0.191, \quad \gamma = 1.80. \quad (2)$$

This has the further advantage that, provided the operator P has the Serber form, it is also reasonably consistent with observed data on the scattering of high-energy neutrons by protons (Christian & Hart 1950).

The fact that it seems possible to obtain agreement with so many observed quantities with an interaction involving only four parameters must be regarded as encouraging. There remains one other quantity which is accurately known which depends on the nucleonic interactions at low energies, and which may be predicted, albeit with difficulty, from an assumed form of these interactions. This is the binding energy of the triton. Before the measurements of Kellogg, Rabi, Ramsey & Zacharias (1940) revealed the finite quadrupole moment of the deuteron and hence the non-central character of nuclear forces, the triton binding energy was one of the most important quantities utilized to determine the parameters in the nucleonic interaction. Calculations were even made for the four-body helium nucleus to check for consistency. These calculations are complicated enough, but the presence of the non-central component in equation (1) makes the calculation very much worse. The non-central character of the interaction was first allowed for in the pioneer work of Gerjuoy & Schwinger (1942). They assumed an interaction of the form of (1) with the function J of the spherical well type of range 2.8×10^{-13} cm. (the original interaction studied by Rarita & Schwinger). With the trial wave function they assumed,

only 39 % of the observed binding energy was found. This was not very surprising, as the trial function was of simple Gaussian type and did not include all the possible spin and angular components (in Gerjuoy & Schwinger's notation (see §2) only the ${}^2S(\tilde{1}\tilde{2}, 3)$ and one part of the 4D components were included). Feshbach & Rarita (1949) tried to improve the result by using Hylleraas's co-ordinates which have been successfully applied to the study of the helium atom (1930) by the variation method as well as to the triton when purely central interaction is assumed (Rarita & Present 1937). They still did not include all the spin and angular components and did not arrive at a substantially better result than did Gerjuoy & Schwinger. It soon became clear from a study of the character of the triton wave function in relation to the iteration process involved in determining the binding energy, using more and more available parameters, that components neglected in earlier calculations, especially ${}^4D(r_{31}) + {}^4D(r_{23})$, are by no means unimportant. In this paper detailed calculations are described in which allowance was made for these terms, the trial wave function including a complete set of angular and spin functions. The radial forms involved were taken as exponentials multiplied by polynomial expansions in the relative co-ordinates. The results obtained show that the full binding energy can be obtained with non-central forces of the usual kind. Apart from this aspect the opportunity has also been taken to examine how strongly the binding energy depends on the nature of the exchange operator P , the difference in the binding energy of ${}^3\text{He}$ and ${}^3\text{H}$, and the magnetic moments of ${}^3\text{H}$ and ${}^3\text{He}$.

2. THEORY AND METHOD OF CALCULATION

Treating protons and neutrons as the two possible states of nucleons, we must have, according to the Pauli exclusion principle, a totally antisymmetrical wave function for the three particles, which can be written as a combination

$$\Psi(123) = \frac{1}{\sqrt{3}} [\psi(\tilde{1}\tilde{2}, 3) + \psi(\tilde{2}\tilde{3}, 1) + \psi(\tilde{3}\tilde{1}, 2)] \quad (3)$$

of wave functions, such as $\psi(\tilde{1}\tilde{2}, 3)$, antisymmetrical with respect to a pair 1, 2 of nucleons. By further factorizing the wave function into two parts describing the spatial and spin co-ordinates and the charge states respectively, we have

$$\psi(\tilde{1}\tilde{2}, 3) = \phi_1(\tilde{1}\tilde{2}, 3)\tau(\overline{1}\overline{2}, 3) + \phi_2(\overline{1}\overline{2}, 3)\tau(\tilde{1}\tilde{2}, 3). \quad (4)$$

We denote here and henceforth a function which is symmetrical with respect to a pair of nucleons (1,2) as $f(\overline{1}\overline{2}, 3)$ and that antisymmetrical as $f(\tilde{1}\tilde{2}, 3)$. The isotopic spin functions, which are simultaneous eigenfunctions of the equations

$$\frac{1}{4}(\sum \tau_i)^2 \tau(123) = T(T+1)\tau(123),$$

$$\frac{1}{2}\sum \tau_i^3 \tau(123) = T^{(3)}\tau(123),$$

are as follows: ${}^2\tau(\tilde{1}\tilde{2}, 3) = \frac{1}{\sqrt{2}} [a(1)b(2)a(3) - b(1)a(2)a(3)], \quad (5a)$

$${}^2\tau(\overline{1}\overline{2}, 3) = \frac{1}{\sqrt{6}} [a(1)b(2)a(3) + b(1)a(2)a(3) - 2a(1)a(2)b(3)], \quad (5b)$$

$${}^4\tau(\overline{1}\overline{2}, 3) = \frac{1}{\sqrt{3}} [a(1)b(2)a(3) + b(1)a(2)a(3) + a(1)a(2)b(3)], \quad (6)$$

for two neutrons and one proton, and

$${}^4\tau(123) = a(1)a(2)a(3) \quad (7)$$

for three neutrons. τ denotes the isotopic spin operator and a, b the two isotopic spin wave functions. There are also similar forms for two protons and one neutron and for three protons.

Under the hypothesis of charge independence of nuclear forces, the operator $(\Sigma\tau_i)^2$ commutes with the total Hamiltonian and hence has τ eigenvalues as good quantum numbers. The ground state of a nucleus will take its lowest possible τ eigenvalue, so the ground state of the triton is a pure charge doublet. On the other hand, the Coulomb force existing in the ${}^3\text{He}$ nucleus will induce some amount of charge quartet in addition to the charge doublet. It is, however, much weaker than the nuclear forces and will not alter the fundamental features substantially.

For application of the variation principle, we usually write the wave function in the individual particle form

$$\Psi(123) = \frac{1}{\sqrt{3}} [\Phi(\tilde{1}\bar{2}, 3)a(1)a(2)b(3) + \Phi(\tilde{2}\bar{3}, 1)a(2)a(3)b(1) + \Phi(\tilde{3}\bar{1}, 2)a(3)a(1)b(2)], \quad (8)$$

but this function will in general correspond to a mixture of charge doublets and charge quartets. It is necessary to have them separated into pure states as follows.

On substitution of the isotopic functions in equations (3) and (4) we have, for a charge doublet,

$$\begin{aligned} \Psi(123) &= \frac{1}{\sqrt{3}} \sum_{\text{cycl.}} \{ \phi_1(\tilde{1}\bar{2}, 3)\tau(\bar{1}\bar{2}, 3) + \phi_2(\bar{1}\bar{2}, 3)\tau(\tilde{1}\bar{2}, 3) \} \\ &= \frac{1}{\sqrt{3}} \sum_{\text{cycl.}} - \frac{1}{\sqrt{6}} \left\{ 2 \left[\phi_1(\tilde{1}\bar{2}, 3) - \frac{1}{\sqrt{3}} \phi_2(\bar{2}\bar{3}, 1) + \frac{1}{\sqrt{3}} \phi_2(\bar{3}\bar{1}, 2) \right] \right. \\ &\quad \left. - \left[\phi_1(\tilde{2}\bar{3}, 1) - \frac{1}{\sqrt{3}} \phi_2(\bar{3}\bar{1}, 2) + \frac{1}{\sqrt{3}} \phi_2(\bar{1}\bar{2}, 3) \right] \right. \\ &\quad \left. - \left[\phi_1(\tilde{3}\bar{1}, 2) - \frac{1}{\sqrt{3}} \phi_2(\bar{1}\bar{2}, 3) + \frac{1}{\sqrt{3}} \phi_2(\bar{2}\bar{3}, 1) \right] \right\} a(1)a(2)b(3), \end{aligned}$$

so that Ψ may be written in the form (8) with

$$\Phi(\tilde{1}\bar{2}, 3) = (2 + P_{13} + P_{23})\phi(\tilde{1}\bar{2}, 3). \quad (9a)$$

Similarly, for a charge quartet,

$$\begin{aligned} \Psi(123) &= \frac{1}{\sqrt{3}} [\phi_1(\tilde{1}\bar{2}, 3) + \phi_1(\tilde{2}\bar{3}, 1) + \phi_1(\tilde{3}\bar{1}, 2)] \tau(123) \\ &= \frac{1}{\sqrt{3}} \sum_{\text{cycl.}} \frac{1}{\sqrt{3}} [\phi_1(\tilde{1}\bar{2}, 3) + \phi_1(\tilde{2}\bar{3}, 1) + \phi_1(\tilde{3}\bar{1}, 2)] a(1)a(2)b(3), \end{aligned}$$

so that, in the form (8)

$$\Phi(\tilde{1}\bar{2}, 3) = (1 - P_{13} - P_{23})\phi(\tilde{1}\bar{2}, 3). \quad (9b)$$

P_{ij} is the operator which interchanges the co-ordinates of particles (1, 2).

For the function ϕ we shall now obtain a set of simultaneous eigenfunctions of J^2, L^2 and S^2 . This is possible since these quantities all commute with each other.

However, it is only for purely central forces that S and L will be conserved. With tensor forces the solution of the Schrödinger equation will be a combination of various components with different L and S but with certain definite J .

The simultaneous eigenfunctions can be set up by solving the equations

$$S^2\phi = s(s+1)\hbar^2\phi,$$

$$L^2\phi = l(l+1)\hbar^2\phi,$$

and

$$J^2\phi = \{L^2 + S^2 + \hbar(\Sigma\sigma_i) \cdot \mathbf{L}\} \phi = j(j+1)\hbar^2\phi.$$

Instead of the analytical form, it is sometimes more convenient to use the operator form. For instance, we have the following even-parity functions with $J = \frac{1}{2}$ and $J_z = +\frac{1}{2}$ for the triton (Gerjuoy & Schwinger 1942):

$$\left. \begin{aligned} {}^2\tilde{S} &= \frac{1}{\sqrt{2}}[\alpha(1)\beta(2)\alpha(3) - \beta(1)\alpha(2)\alpha(3)] = \chi, \\ {}^2\bar{S} &= \sigma_1 \cdot \sigma_3 \chi, \\ {}^2\tilde{P} &= i\sigma_3 \cdot (\mathbf{r}_{23} \times \mathbf{r}_{31}) \chi, \\ {}^2\bar{P} &= (\sigma_1 \cdot \sigma_3) {}^2\tilde{P} = [i\sigma_1 \cdot (\mathbf{r}_{23} \times \mathbf{r}_{31}) + (\sigma_1 \times \sigma_3) \cdot (\mathbf{r}_{23} \times \mathbf{r}_{31})] \chi, \\ {}^4P &= [i\sigma_1 \cdot (\mathbf{r}_{23} \times \mathbf{r}_{31}) - \frac{1}{2}(\sigma_1 \times \sigma_3) \cdot (\mathbf{r}_{23} \times \mathbf{r}_{31})] \chi, \\ {}^4D(r_{12}) &= [3(\sigma_1 \cdot \mathbf{r}_{12})(\sigma_3 \cdot \mathbf{r}_{12}) - r_{12}^2(\sigma_1 \cdot \sigma_3)] \chi, \\ {}^4D(r_{31}) \pm {}^4D(r_{23}) &= [3(\sigma_1 \cdot \mathbf{r}_{31})(\sigma_3 \cdot \mathbf{r}_{31}) - r_{31}^2(\sigma_1 \cdot \sigma_3) \pm 3(\sigma_1 \cdot \mathbf{r}_{23})(\sigma_3 \cdot \mathbf{r}_{23}) \mp r_{23}^2(\sigma_1 \cdot \sigma_3)] \chi, \end{aligned} \right\} \quad (10)$$

where σ is the spin operator, α, β the two spin wave functions and $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ the relative co-ordinate of the particle j from i . They all satisfy the equation

$$\frac{1}{2}\Sigma\nabla_i^2\phi = \frac{1}{2}[(\nabla_{12} - \nabla_{23})^2 + (\nabla_{23} - \nabla_{31})^2 + (\nabla_{31} - \nabla_{12})^2]\phi = 0. \quad (11)$$

We now choose a suitable trial form for the function $\phi(\tilde{1}2, 3)$ and thence for $\Phi(\tilde{1}2, 3)$. These functions must be proper functions involving linear combinations of the forms (10). We take

$$\begin{aligned} \phi(12, 3) &= \exp\left\{-\frac{1}{2}\mu(r_{12} + r_{23} + r_{31})\right\} [G_1 {}^2\tilde{S} + G_2 {}^2\bar{S}] \\ &\quad + \exp\left\{-\frac{1}{2}\nu(r_{12} + r_{23} + r_{31})\right\} [G_3 {}^2\tilde{P} + G_4 {}^2\bar{P} + G_5 {}^4P] \\ &\quad + \exp\left\{-\frac{1}{2}\omega(r_{12} + r_{23} + r_{31})\right\} [G_6 {}^4D(\mathbf{r}_{12}) + G_7\{{}^4D(\mathbf{r}_{31}) + {}^4D(\mathbf{r}_{32})\} \\ &\quad \quad \quad + G_8\{{}^4D(\mathbf{r}_{31}) - {}^4D(\mathbf{r}_{32})\}], \end{aligned} \quad (12)$$

where

$$\begin{aligned} G_i &= d_{000}^{(i)} + d_{100}^{(i)}r_{12} + d_{010}^{(i)}(r_{31} + r_{23}) + d_{001}^{(i)}(r_{31} - r_{23}) \\ &\quad + d_{200}^{(i)}r_{12}^2 + d_{110}^{(i)}r_{12}(r_{31} + r_{23}) + d_{020}^{(i)}(r_{31} + r_{23})^2 + d_{002}^{(i)}(r_{31} - r_{23})^2 \\ &\quad + d_{101}^{(i)}r_{12}(r_{31} - r_{23}) + d_{011}^{(i)}(r_{31} + r_{23})(r_{31} - r_{23}) + \dots, \end{aligned} \quad (13)$$

an expansion in powers of $r_{12}, r_{31} \pm r_{23}$, G_1, G_4, G_8 being functions symmetrical and G_2, G_3, G_5, G_6, G_7 antisymmetrical with respect to (1, 2).

The function $\Phi(\tilde{1}\tilde{2}, 3)$ derived from $\phi(\tilde{1}\tilde{2}, 3)$ by the relations (9) can be written in a similar form:

$$\begin{aligned} \Phi(\tilde{1}\tilde{2}, 3) = & \exp\left\{-\frac{1}{2}\mu(r_{12}+r_{23}+r_{31})\right\} [F_1 {}^2\tilde{S} + F_2 {}^2\bar{S}] \\ & + \exp\left\{-\frac{1}{2}\nu(r_{12}+r_{23}+r_{31})\right\} [F_3 {}^2\tilde{P} + F_4 {}^2\bar{P} + F_5 {}^4P] \\ & + \exp\left\{-\frac{1}{2}\omega(r_{12}+r_{23}+r_{31})\right\} [F_6 {}^4D(\mathbf{r}_{12}) + F_7 \{ {}^4D(\mathbf{r}_{31}) + {}^4D(\mathbf{r}_{23}) \} \\ & + F_8 \{ {}^4D(\mathbf{r}_{31}) - {}^4D(\mathbf{r}_{23}) \}], \quad (14) \end{aligned}$$

with F_i a similar polynomial to G_i in which the coefficients $d_{lmn}^{(i)}$ are replaced by $c_{lmn}^{(i)}$.

To satisfy (9) relations must exist between the c 's and the d 's. To obtain these we substitute for ϕ in (9) and use the relations

$$\begin{aligned} P_{13} {}^2\tilde{S} &= \frac{1}{2}({}^2\tilde{S} + {}^2\bar{S}), \quad P_{13} {}^2\bar{S} = \frac{1}{2}(3 {}^2\tilde{S} - {}^2\bar{S}), \\ P_{13} {}^2\tilde{P} &= -\frac{1}{2}({}^2\tilde{P} + {}^2\bar{P}), \quad P_{13} {}^2\bar{P} = -\frac{1}{2}(3 {}^2\tilde{P} - {}^2\bar{P}), \quad P_{13} {}^4P = -{}^4P, \\ P_{13} {}^4D(\mathbf{r}_{12}) &= {}^4D(\mathbf{r}_{23}), \quad P_{13} \{ {}^4D(\mathbf{r}_{31}) \pm {}^4D(\mathbf{r}_{23}) \} = \pm {}^4D(\mathbf{r}_{12}) + {}^4D(\mathbf{r}_{31}), \end{aligned}$$

together with the similar set for P_{23} . This gives, for a charge doublet,

$$\left. \begin{aligned} F_1 &= (2 + \frac{1}{2}P_{13} + \frac{1}{2}P_{23}) G_1 + \frac{3}{2}(P_{13} - P_{23}) G_2, \\ F_2 &= \frac{1}{2}(P_{13} - P_{23}) G_1 + (2 - \frac{1}{2}P_{13} - \frac{1}{2}P_{23}) G_2, \\ F_3 &= (2 - \frac{1}{2}P_{13} - \frac{1}{2}P_{23}) G_3 - \frac{3}{2}(P_{13} - P_{23}) G_4, \\ F_4 &= -\frac{1}{2}(P_{13} - P_{23}) G_3 + (2 + \frac{1}{2}P_{13} + \frac{1}{2}P_{23}) G_4, \\ F_5 &= (2 - P_{13} - P_{23}) G_5, \\ F_6 &= 2G_6 + (P_{13} + P_{23}) G_7 - (P_{13} - P_{23}) G_8, \\ F_7 &= \frac{1}{2}(P_{13} + P_{23}) G_6 + (2 + \frac{1}{2}P_{13} + \frac{1}{2}P_{23}) G_7 + \frac{1}{2}(P_{13} - P_{23}) G_8, \\ F_8 &= -\frac{1}{2}(P_{13} - P_{23}) G_6 + \frac{1}{2}(P_{13} - P_{23}) G_7 + (2 + \frac{1}{2}P_{13} + \frac{1}{2}P_{23}) G_8, \end{aligned} \right\} \quad (15)$$

and a similar set of expressions for a charge quartet. Equating the coefficients of the corresponding terms in the polynomial we have, for instance, for the charge doublet of the 2S component,

$$\begin{aligned} c_{000}^{(1)} &= 3d_{000}^{(1)}, \\ c_{100}^{(1)} &= 2d_{100}^{(1)} + d_{010}^{(1)} - 3d_{001}^{(2)}, \\ c_{010}^{(1)} &= \frac{1}{2}d_{100}^{(1)} + \frac{5}{2}d_{010}^{(1)} + \frac{3}{2}d_{001}^{(2)}, \\ c_{001}^{(2)} &= -\frac{1}{2}d_{100}^{(1)} + \frac{1}{2}d_{010}^{(1)} + \frac{3}{2}d_{001}^{(2)}. \end{aligned}$$

The only term of zero order is independent, but for the first-order terms there exists a relation

$$c_{100}^{(1)} - c_{010}^{(1)} + 3c_{001}^{(2)} = 0,$$

because the functional determinant vanishes. We have then the following relations between the coefficients c :

For a pure charge doublet:

2S No condition for zero-order term,

$$\begin{aligned} c_{100}^{(1)} - c_{010}^{(1)} + 3c_{001}^{(2)} &= 0, \\ c_{200}^{(1)} - c_{020}^{(1)} - c_{002}^{(1)} + 3c_{011}^{(2)} &= 0, \\ c_{110}^{(1)} - 2c_{020}^{(1)} + 2c_{002}^{(1)} + 3c_{101}^{(2)} &= 0. \end{aligned}$$

2P

No condition for zero-order term,
$$c_{001}^{(3)} - c_{100}^{(4)} + c_{010}^{(4)} = 0,$$
$$c_{101}^{(3)} - c_{110}^{(4)} + 2c_{020}^{(4)} - 2c_{002}^{(4)} = 0,$$
$$c_{011}^{(3)} - c_{200}^{(4)} + c_{020}^{(4)} + c_{002}^{(4)} = 0.$$

4P

$c_{000}^{(5)} = 0,$
$$c_{100}^{(5)} + 2c_{010}^{(5)} = 0,$$
$$c_{200}^{(5)} + 2c_{020}^{(5)} + 2c_{002}^{(5)} = 0,$$
$$c_{110}^{(5)} + c_{020}^{(5)} - c_{002}^{(5)} = 0.$$

4D

No condition for zero-order term,
$$c_{001}^{(6)} - c_{001}^{(7)} - c_{100}^{(8)} + c_{010}^{(8)} = 0,$$
$$c_{101}^{(6)} - c_{101}^{(7)} - c_{110}^{(8)} + 2c_{020}^{(8)} - 2c_{002}^{(8)} = 0,$$
$$c_{011}^{(6)} - c_{011}^{(7)} - c_{200}^{(8)} + c_{020}^{(8)} + c_{002}^{(8)} = 0.$$

A similar set of relations exist for a pure charge quartet. These constitute two sets of orthogonal functions and therefore afford a very useful means for checking the calculations of all kinds of matrix elements.

FIGURE 1. Coupling diagram.

The various components are coupled together as shown in figure 1. With Wigner approximation (1937) the ground state of the triton corresponds to a supermultiplet $(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2})$. We can therefore regard it as consisting of mainly the ${}^2\tilde{S}$ component and take it as an initial approximate solution for iteration purposes. Owing to the interaction between pairs (2, 3) and (3, 1) whereby

$$[J_{23}\boldsymbol{\sigma}_2\cdot\boldsymbol{\sigma}_3 + J_{31}\boldsymbol{\sigma}_3\cdot\boldsymbol{\sigma}_1] {}^2\tilde{S} = [J_{31} - J_{23}] {}^2\bar{S},$$
$$[J_{23}S_{23}(\mathbf{r}_{23}) + J_{31}S_{31}(\mathbf{r}_{31})] {}^2\tilde{S} = \frac{1}{2}\left(\frac{J_{31}}{r_{31}^2} + \frac{J_{23}}{r_{23}^2}\right) ({}^4D(\mathbf{r}_{31}) - {}^4D(\mathbf{r}_{23}))$$
$$+ \frac{1}{2}\left(\frac{J_{31}}{r_{31}^2} - \frac{J_{23}}{r_{23}^2}\right) ({}^4D(\mathbf{r}_{31}) + {}^4D(\mathbf{r}_{23})),$$

${}^2\bar{S}$ and ${}^4D(\mathbf{r}_{31}) \pm {}^4D(\mathbf{r}_{23})$ will appear as first-order terms. The component

$${}^4D(\mathbf{r}_{31}) + {}^4D(\mathbf{r}_{23}),$$

which has been neglected in previous calculations, in fact plays a very important role in the triton wave function.

The variation principle requires the integral

$$I = \int \Psi^*(123) (H - E) \Psi(123) d\tau = 0 \quad (16)$$

to be stationary. This integral consists of three parts,

$$I = I(\text{K.E.}) + I(\text{P.E.}) - I(\text{B.E.}), \quad (17)$$

where

$$I(\text{K.E.}) = -\frac{\hbar^2}{2M} \int \Phi^*(\tilde{1}2, 3) (\Sigma_i \nabla_i^2) \Phi(\tilde{1}2, 3) d\tau, \quad (18)$$

$$I(\text{P.E.}) = \int \Phi^*(\tilde{1}2, 3) (\mathcal{V}_{12} + \mathcal{V}_{23} + \mathcal{V}_{31}) \Phi(\tilde{1}2, 3) d\tau \quad (\text{ordinary}), \quad (19a)$$

$$= \int \Phi^*(\tilde{1}2, 3) [\mathcal{V}_{12} \Phi(\tilde{1}2, 3) + \mathcal{V}_{23} \Phi(\tilde{3}1, 2) + \mathcal{V}_{31} \Phi(\tilde{2}3, 1)] d\tau \quad (\text{exchange}), \quad (19b)$$

$$I(\text{B.E.}) = E \int \Phi^*(\tilde{1}2, 3) \Phi(\tilde{1}2, 3) d\tau. \quad (20)$$

The Coulomb interaction in the ^3He nucleus contributes another term to the integral

$$I(\text{C.E.}) = e^2 \int \Phi^*(\tilde{1}2, 3) r_{12}^{-1} \Phi(\tilde{1}2, 3) d\tau.$$

The integrals involved are given in the appendix. Those arising from the interaction between pairs (2, 3) and (3, 1) can be changed cyclically into those involving (1, 2); thus

$$\begin{aligned} \int \Phi^*(\tilde{1}2, 3) (\mathcal{V}_{23} + \mathcal{V}_{31}) \Phi(\tilde{1}2, 3) d\tau &= 2 \int \Phi^*(\tilde{1}2, 3) \mathcal{V}_{12} \Phi(\tilde{1}2, 3) d\tau, \\ \int \Phi^*(\tilde{1}2, 3) [\mathcal{V}_{23} \Phi(\tilde{3}1, 2) + \mathcal{V}_{31} \Phi(\tilde{2}3, 1)] d\tau &= 2 \int \Phi^*(\tilde{3}1, 2) \mathcal{V}_{12} \Phi(\tilde{2}3, 1) d\tau. \end{aligned}$$

With the trial wave function used $\Phi(\tilde{2}3, 1)$ and $\Phi(\tilde{3}1, 2)$ are simply another combination of the components bearing definite relations between the coefficients so that only those integrals for $\mathcal{V}_{12}$ are needed.

The variation calculation is worked out as usual by solving a characteristic determinant, the exponents μ, ν, ω having been fixed at some suitable values found by preliminary calculations. The values of μ, ν, ω finally used were 1.5, 3, 3 respectively. Previous to the actual evaluation of the determinant, using the preliminary calculations as a guide, a value for the binding energy must be first assumed and then corrected by relaxation. Lastly, it should be pointed out that the heavy numerical work involved in this problem makes imperative the use of extremely systematic methods with carefully planned checks in all the steps.

3. RESULTS AND DISCUSSION

A summary of the results obtained is given in table 1. Most detailed calculations were carried out for an interaction of Yukawa shape of range 1.36×10^{-13} cm. This is consistent with the other low-energy data. The results given in table 1 show that

it probably gives nearly the correct binding energy of the triton. Thus the effect of including second-order terms in $\bar{F}_1, \bar{F}_2, \bar{F}_6$ (see (15)) is to increase the calculated value by 1.7 to 2.2 MeV, and it is then between 1.6 and 1.3 MeV below the correct value. The slowness of convergence suggests that the higher approximations are quite likely to contribute enough to bridge this gap. At any rate it is probable that there is no very large discrepancy such as appeared likely from the earlier calculations. On the other hand, the slowness of convergence will make it very difficult to use the triton binding energy as an accurate test of any non-central interaction.

TABLE 1. BINDING ENERGY OF THE TRITON

shape of the well	r_c (10^{-13} cm.)	r_N/r_c	exchange nature	$-E(^3\text{H})$ (MeV)	$-E(^3\text{He})$ (MeV)	trial wave function
Yukawa	1.36*	1	exchange	7.0	—	including second-order terms in $\bar{F}_1, \bar{F}_2, \bar{F}_6, \bar{F}_7$ and $\bar{F}_8$ and first-order terms in $\bar{F}_3, \bar{F}_4$ and $\bar{F}_5$; $\mu = 1.5, \nu = \omega = 3$
			ordinary	6.7	—	
Yukawa	1.36	1	exchange	5.28	4.26	including all the first-order terms, $\mu = 1.5, \nu = \omega = 3$
			ordinary	4.52	3.50	
	1.74†	1	ordinary	6.5	—	
	2.18†	1	ordinary	6.4	—	
Yukawa	3.05†	1	ordinary	5.4	—	including first order terms for 2S -waves
	1.31	∞	ordinary	12.16	—	
	1.74	∞	ordinary	9.00	—	
	2.18	∞	ordinary	7.43	—	
Yukawa	3.05	∞	ordinary	5.44	—	including terms: $c_{000}^{(1)}, c_{001}^{(7)}, c_{000}^{(8)}$ and $c_{100}^{(8)}$
	0.80‡	2	ordinary	10.2	—	
spherical	2.80§	1	ordinary	4.0	—	including terms: $c_{000}^{(1)}, c_{001}^{(7)}$ and $c_{000}^{(8)}$.

* K. N. Hsu (unpublished): $a = 1.14, \gamma = 1.80, g = -0.191$.
† T. M. Hu & H. S. W. Massey (1949).
‡ H. N. Yadav (1949).
§ W. Rarita & J. Schwinger (1941a).

To examine the effect of the tensor forces more closely, it is convenient to examine the variation of the calculated binding energy with range of the central forces with and without inclusion of tensor forces. This may be done by reference to figure 2. Curve I is obtained from the calculations for purely central forces, curve II from the calculations of this paper in which the central and non-central forces are taken to have the same range. In all cases the interactions are consistent with the binding energy of the deuteron, the low-velocity neutron-proton scattering cross-section and the charge independence of the forces. The non-central interactions all give the correct quadrupole moment of the deuteron and are not inconsistent with the rather indefinite evidence from its magnetic moment. It will be seen that the presence of the non-central component reduces the binding energy for a given range of interactions by an amount increasing with decrease of range. An intermediate case between the sets of interactions concerned in curves I and II is represented by the point A in the diagram. This was calculated for an interaction in which the range of the central forces r_c was taken to be half of that, r_N , for the tensor forces with $r_c = 0.80 \times 10^{-13}$ cm., the other constants being adjusted to agree with the same data

as the interactions of curve II (see table 1). As expected, this gives, for the same values of r_c , higher binding energy than for $r_N/r_c = 1$ (curve II), and a lower than for $r_N/r_c \rightarrow \infty$ (curve I). The slowness of convergence of the variational method increases with decrease of r_c for any particular value of r_N/r_c and with decrease of r_N/r_c for any particular r_c . This makes it difficult to say whether the binding energy actually decreases monotonically with increase of r_c for $r_N/r_c = 1$ or behaves as indicated in curve II of figure 2, which only refers to the results of first-order calculations. Thus the point B of figure 2 calculated, including second-order terms, is not inconsistent with monotonic variation.

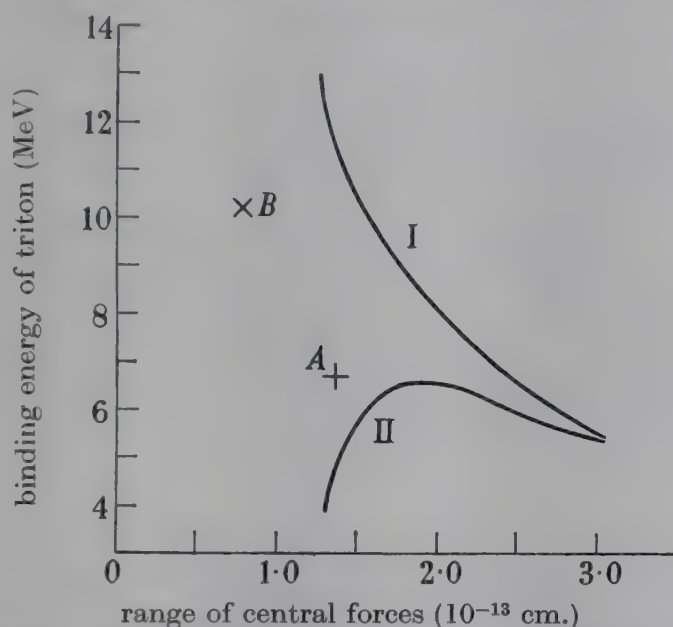


FIGURE 2. The variation of the triton binding energy with the range of nuclear forces assuming Yukawa well interactions with no exchange operator. Curve I, $r_N/r_c = \infty$; curve II, $r_N/r_c = 1$ (including first-order terms only). $A+$, $r_N/r_c = 1$ (including second-order terms). $B \times$, $r_N/r_c = 2$ (including terms $c_{000}^{(1)}$, $c_{001}^{(7)}$, $c_{000}^{(8)}$ and $c_{100}^{(8)}$).

The effect of the exchange operator P was investigated by considering only the most important terms of the ${}^2\tilde{S}$ wave ($\chi \exp \{ -\frac{1}{2}\mu(r_{12} + r_{23} + r_{31}) \}$) and the 4D waves. The matrix element $\langle {}^2\tilde{S} | V | {}^2\tilde{S} \rangle$ depends only on the intensities of the 1S and 3S interaction for the deuteron and is independent of the assumed exchange properties. The coupling matrix elements $\langle {}^4D | V | {}^2\tilde{S} \rangle$ are also independent of the exchange operator. Even for the matrix elements $\langle {}^4D | V | {}^4D \rangle$ which do depend on this operator, the difference between the values given with and without its inclusion are so small compared with the kinetic energy terms that no marked effect is apparent in the calculated binding energy. This is apparent from the results given in table 1 for forces of range 1.36×10^{-13} cm.

The ${}^3\text{He}$ problem differs from that of the triton in the presence of the Coulomb force, which, being charge dependent, introduces a small admixture of charge quartet waves. This seems to be very ineffective in modifying the binding energy. Calculations carried out, including all first-order terms for the range of 1.36×10^{-13} cm., give a difference between the binding energies of ${}^3\text{He}$ and ${}^3\text{H}$ of exactly the magnitude obtained by treating the Coulomb field as a perturbation of the triton system. The assumption that electrostatic repulsion can account for all the difference in binding energy between ${}^3\text{H}$ and ${}^3\text{He}$ is therefore justified.

We have also investigated the magnetic moments of ^3H and ^3He on the assumptions that the intrinsic magnetic moments of the nucleons are additive and the relativistic effects are negligible (Sachs & Schwinger 1946; Sachs 1947). The cross-terms contribute little to the magnetic moment due to the fact that the ground state consists of only a very small percentage of the P -waves, and that it assumes the lowest possible states which give almost vanishing first derivatives. It is impossible to improve the results even qualitatively. Either the exchange magnetic moment should be assumed or the spin-orbit coupling force introduced.

5. APPENDIX

(a) The normalizing factor

The integrands arising from the spin and angular parts of the wave function after summing over spins are as follows:

$$\left. \begin{aligned} \langle {}^2\tilde{S} | {}^2\tilde{S} \rangle &= 1, & \langle {}^2\bar{S} | {}^2\bar{S} \rangle &= 3, & \langle {}^2\tilde{P} | {}^2\tilde{P} \rangle &= (\mathbf{r}_{23} \times \mathbf{r}_{31})^2, \\ \langle {}^2\bar{P} | {}^2\bar{P} \rangle &= 3(\mathbf{r}_{23} \times \mathbf{r}_{31})^2, & \langle {}^4P | {}^4P \rangle &= \frac{3}{2}(\mathbf{r}_{23} \times \mathbf{r}_{31})^2, \\ \langle {}^4D(\mathbf{r}_{12}) | {}^4D(\mathbf{r}_{12}) \rangle &= 6r_{12}^4, & \langle {}^4D(\mathbf{r}_{12}) | {}^4D(\mathbf{r}_{23}) \rangle &= 9(\mathbf{r}_{12} \cdot \mathbf{r}_{23})^2 - 3r_{12}^2 r_{23}^2, \\ \langle {}^4D(\mathbf{r}_{12}) | {}^4D(\mathbf{r}_{31}) \rangle &= 9(\mathbf{r}_{12} \cdot \mathbf{r}_{31})^2 - 3r_{12}^2 r_{31}^2, & \langle {}^4D(\mathbf{r}_{23}) | {}^4D(\mathbf{r}_{23}) \rangle &= 6r_{23}^4, \\ \langle {}^4D(\mathbf{r}_{23}) | {}^4D(\mathbf{r}_{31}) \rangle &= 9(\mathbf{r}_{23} \cdot \mathbf{r}_{31})^2 - 3r_{23}^2 r_{31}^2, & \langle {}^4D(\mathbf{r}_{31}) | {}^4D(\mathbf{r}_{31}) \rangle &= 6r_{31}^4. \end{aligned} \right\} \quad (22)$$

(b) The potential energy

For the Coulomb and central forces the integrands arising from the spin and angular part of the wave function are similar to those occurring in the normalizing

factor. For tensor forces we have, writing $S_{12} = \frac{3\boldsymbol{\sigma}_1 \cdot \mathbf{r}_{12} \boldsymbol{\sigma}_2 \cdot \mathbf{r}_{12}}{r_{12}^2} - \boldsymbol{\sigma}_1 \cdot \boldsymbol{\sigma}_2$,

$$\left. \begin{aligned} \langle {}^2\bar{S} | S_{12} | {}^4D(\mathbf{r}_{12}) \rangle &= -12r_{12}^2, \\ \langle {}^2\bar{S} | S_{12} | {}^4D(\mathbf{r}_{23}) \rangle &= 6r_{23}^2 - 18(\mathbf{r}_{12} \cdot \mathbf{r}_{23})^2/r_{12}^2, \\ \langle {}^2\bar{S} | S_{12} | {}^4D(\mathbf{r}_{31}) \rangle &= 6r_{31}^2 - 18(\mathbf{r}_{12} \cdot \mathbf{r}_{31})^2/r_{12}^2, \\ \langle {}^2\bar{P} | S_{12} | {}^4P \rangle &= 3(\mathbf{r}_{23} \times \mathbf{r}_{31})^2, \\ \langle {}^2\bar{P} | S_{12} | {}^4D(\mathbf{r}_{12}) \rangle &= -18(\mathbf{r}_{12} \cdot \mathbf{r}_{23})(\mathbf{r}_{23} \times \mathbf{r}_{31})^2/r_{12}r_{23}, \\ \langle {}^2\bar{P} | S_{12} | {}^4D(\mathbf{r}_{31}) \rangle &= 18(\mathbf{r}_{12} \cdot \mathbf{r}_{31})(\mathbf{r}_{23} \times \mathbf{r}_{31})^2/r_{12}r_{31}, \\ \langle {}^4P | S_{12} | {}^4P \rangle &= \frac{3}{2}(\mathbf{r}_{23} \times \mathbf{r}_{31})^2, \\ \langle {}^4P | S_{12} | {}^4D(\mathbf{r}_{23}) \rangle &= 9(\mathbf{r}_{12} \cdot \mathbf{r}_{23})(\mathbf{r}_{23} \times \mathbf{r}_{31})^2/r_{12}r_{23}, \\ \langle {}^4P | S_{12} | {}^4D(\mathbf{r}_{31}) \rangle &= -9(\mathbf{r}_{12} \cdot \mathbf{r}_{31})(\mathbf{r}_{23} \times \mathbf{r}_{31})^2/r_{12}r_{31}, \\ \langle {}^4D(\mathbf{r}_{12}) | S_{12} | {}^4D(\mathbf{r}_{12}) \rangle &= -12r_{12}^4, \\ \langle {}^4D(\mathbf{r}_{12}) | S_{12} | {}^4D(\mathbf{r}_{23}) \rangle &= -18(\mathbf{r}_{12} \cdot \mathbf{r}_{23})^2 + 6r_{12}^2 r_{23}^2, \\ \langle {}^4D(\mathbf{r}_{12}) | S_{12} | {}^4D(\mathbf{r}_{31}) \rangle &= -18(\mathbf{r}_{12} \cdot \mathbf{r}_{31})^2 + 6r_{12}^2 r_{31}^2, \\ \langle {}^4D(\mathbf{r}_{23}) | S_{12} | {}^4D(\mathbf{r}_{23}) \rangle &= 6r_{23}^4 - 18(\mathbf{r}_{12} \cdot \mathbf{r}_{23})^2 r_{23}^2/r_{12}^2, \\ \langle {}^4D(\mathbf{r}_{23}) | S_{12} | {}^4D(\mathbf{r}_{31}) \rangle &= r_{12}^{-2} \{ -54(\mathbf{r}_{12} \cdot \mathbf{r}_{23})(\mathbf{r}_{23} \cdot \mathbf{r}_{31})(\mathbf{r}_{31} \cdot \mathbf{r}_{12}) + 18(\mathbf{r}_{12} \cdot \mathbf{r}_{23})^2 r_{31}^2 \\ &\quad + 18(\mathbf{r}_{12} \cdot \mathbf{r}_{31})^2 r_{23}^2 \} + 18(\mathbf{r}_{23} \cdot \mathbf{r}_{31})^2 - 12, \\ \langle {}^4D(\mathbf{r}_{31}) | S_{12} | {}^4D(\mathbf{r}_{31}) \rangle &= 6r_{31}^4 - 18(\mathbf{r}_{12} \cdot \mathbf{r}_{31})^2 r_{31}^2/r_{12}^2. \end{aligned} \right\} \quad (23)$$

(c) The kinetic energy

If $\phi = fg$ and $\Sigma_i \nabla_i^2 g = 0$, then

$$-\frac{1}{2} \int \phi^* (\Sigma_i \nabla_i^2 \phi) d\tau = \int g^2 P f d\tau,$$

where

$$P f = \left(\frac{\partial f}{\partial r_{12}} \right)^2 + \left(\frac{\partial f}{\partial r_{23}} \right)^2 + \left(\frac{\partial f}{\partial r_{31}} \right)^2 - \frac{\mathbf{r}_{12} \cdot \mathbf{r}_{23}}{r_{12} r_{23}} \frac{\partial f}{\partial r_{12}} \frac{\partial f}{\partial r_{23}} - \frac{\mathbf{r}_{23} \cdot \mathbf{r}_{31}}{r_{23} r_{31}} \frac{\partial f}{\partial r_{23}} \frac{\partial f}{\partial r_{31}} - \frac{\mathbf{r}_{31} \cdot \mathbf{r}_{12}}{r_{31} r_{12}} \frac{\partial f}{\partial r_{31}} \frac{\partial f}{\partial r_{12}}. \quad (24)$$

For cross-terms between the 4D -components we have

$$\begin{aligned} & -\frac{1}{2} \int \phi^* (\Sigma_i \nabla_i^2) \{f {}^4D(\mathbf{r}_{12})\} d\tau \\ & = - \int \phi^* \left\{ \left[P f + \frac{6}{r_{12}} \frac{\partial f}{\partial r_{12}} + \frac{3}{r_{23}} \frac{\partial f}{\partial r_{23}} + \frac{3}{r_{31}} \frac{\partial f}{\partial r_{31}} \right] {}^4D(\mathbf{r}_{12}) \right. \\ & \quad \left. + \left[-\frac{1}{r_{23}} \frac{\partial f}{\partial r_{23}} + \frac{1}{r_{31}} \frac{\partial f}{\partial r_{31}} \right] [{}^4D(\mathbf{r}_{31}) - {}^4D(\mathbf{r}_{23})] \right\} d\tau, \end{aligned} \quad (25a)$$

$$\begin{aligned} & -\frac{1}{2} \int \phi^* (\Sigma_i \nabla_i^2) \{f [{}^4D(\mathbf{r}_{31}) + {}^4D(\mathbf{r}_{23})]\} d\tau \\ & = - \int \phi^* \left\{ \left[P f + \frac{2}{r_{12}} \frac{\partial f}{\partial r_{12}} + \frac{5}{r_{23}} \frac{\partial f}{\partial r_{23}} + \frac{5}{r_{31}} \frac{\partial f}{\partial r_{31}} \right] [{}^4D(\mathbf{r}_{31}) + {}^4D(\mathbf{r}_{23})] \right. \\ & \quad + \left[\frac{2}{r_{12}} \frac{\partial f}{\partial r_{12}} - \frac{1}{r_{23}} \frac{\partial f}{\partial r_{23}} - \frac{1}{r_{31}} \frac{\partial f}{\partial r_{31}} \right] {}^4D(\mathbf{r}_{12}) \\ & \quad \left. + \left[-\frac{2}{r_{23}} \frac{\partial f}{\partial r_{23}} + \frac{2}{r_{31}} \frac{\partial f}{\partial r_{31}} \right] [{}^4D(\mathbf{r}_{31}) - {}^4D(\mathbf{r}_{23})] \right\} d\tau, \end{aligned} \quad (25b)$$

$$\begin{aligned} & -\frac{1}{2} \int \phi^* (\Sigma_i \nabla_i^2) \{f [{}^4D(\mathbf{r}_{31}) - {}^4D(\mathbf{r}_{23})]\} d\tau \\ & = - \int \phi^* \left\{ \left[P f + \frac{4}{r_{12}} \frac{\partial f}{\partial r_{12}} + \frac{4}{r_{23}} \frac{\partial f}{\partial r_{23}} + \frac{4}{r_{31}} \frac{\partial f}{\partial r_{31}} \right] [{}^4D(\mathbf{r}_{31}) - {}^4D(\mathbf{r}_{23})] \right. \\ & \quad \left. + \left[-\frac{1}{r_{23}} \frac{\partial f}{\partial r_{23}} + \frac{1}{r_{31}} \frac{\partial f}{\partial r_{31}} \right] [{}^4D(\mathbf{r}_{12}) + {}^4D(\mathbf{r}_{23}) + {}^4D(\mathbf{r}_{31})] \right\} d\tau. \end{aligned} \quad (25c)$$

It is our pleasure to thank Professor H. S. W. Massey, F.R.S., and Dr E. H. S. Burhop for their interest in this work. Thanks are due to Dr H. N. Yadav for providing us with interaction constants in advance of publication.

REFERENCES

- Barnes, G. A., Stafford, G. H. & Wilkinson, D. H. 1950 *Nature*, **165**, 69.
 Bethe, H. 1949 *Phys. Rev.* **76**, 38.
 Blatt, J. M. & Jackson, J. D. 1949 *Phys. Rev.* **76**, 18.
 Castillejo, L. 1951 In the Press.
 Christian, R. S. & Hart, E. W. 1950 *Phys. Rev.* **77**, 441.
 Feshbach, H. & Rarita, W. 1949 *Phys. Rev.* **75**, 1384.
 Gerjuoy, E. & Schwinger, J. 1942 *Phys. Rev.* **61**, 138.
 Hu, T. M. & Massey, H. S. W. 1949 *Proc. Roy. Soc. A*, **196**, 135.
 Hughes, D. J., Burgoyne, M. T. & Ringo, G. K. 1950 *Phys. Rev.* **77**, 291.
 Hylleraas, E. A. 1930 *Z. Phys., Lpz.*, **65**, 209.

- Jackson, J. D. & Blatt, J. M. 1950 *Rev. Mod. Phys.* **22**, 77.
 Kellogg, J. M. B., Rabi, I. I., Ramsey, N. F. & Zacharias, J. R. 1940 *Phys. Rev.* **57**, 677.
 Rarita, W. & Present, R. D. 1937 *Phys. Rev.* **51**, 788.
 Rarita, W. & Schwinger, J. 1941a *Phys. Rev.* **59**, 436.
 Rarita, W. & Schwinger, J. 1941b *Phys. Rev.* **59**, 556.
 Sachs, R. 1947 *Phys. Rev.* **72**, 312.
 Sachs, R. & Schwinger, J. 1946 *Phys. Rev.* **70**, 41.
 Wigner, E. 1937 *Phys. Rev.* **51**, 106.
 Yadav, H. N. 1949 *Phys. Rev.* **6**, 1720.

Experiments on the reaction $T(d, n) {}^4\text{He}$

I. The 90° cross-section

BY D. L. ALLAN AND M. J. POOLE

Atomic Energy Research Establishment, Harwell, Berks

(Communicated by Sir John Cockcroft, F.R.S.—Received 24 May 1950—
 Revised 19 July 1950)

The 90° cross-section of the reaction ${}^3_1\text{H}(d, n) {}^4_2\text{He}$

has been investigated over the energy range 100 to 200 keV (energy of bombarding triton) using the 200 keV accelerating set of the establishment. Two methods have been used. As a preliminary experiment the yield of alpha-particles from a thick heavy-ice target was measured per unit charge of incident beam, as a function of deuteron energy, and the variation of cross-section deduced from the gradient of this excitation curve and the range-energy relation for tritons in heavy water. Secondly, a comparison was made between the yield of alpha-particles from the D-T reaction and the yield of protons from the D-D reaction when a beam containing both deuterons and tritons was passed through a heavy-water vapour target. (The energy loss in this target was calculated as only a few hundred electron volts.) To do this a simultaneous observation was made of the protons and alpha-particles using the same counter.

The values obtained for the cross-section have been compared with the resonance formulae given by Bretscher & French (1949) and by Taschek, Everhart, Gittings, Hemmendinger & Jarvis (1948) and have been found to be in disagreement with formulae of this type. From considerations of the absolute magnitude of the cross-section it has been deduced that no conventional theory postulating reaction at a distance equal to the sum of the nuclear radii (cf. Konopinski & Teller 1948) will be able to explain this reaction. The evidence for a low-energy resonance (Allan & Poole 1949) is thought to be inconclusive.

INTRODUCTION

The T-D reaction was first investigated by Baker, Holloway, King & Schreiber (1948), who found a peak in the cross-section at 0.32 MeV triton energy, with a value of 2.8 barns at this energy. Subsequent investigations by Bretscher & French (1949) in the energy region 15 to 128 keV showed the cross-section to be a function of energy rising far more rapidly than would be expected from a simple Gamow theory. On the assumption that the cross-section could be fitted by a Gamow function modified by a resonance term of the form

$$\sigma(E) = \frac{A}{E} \frac{1}{(E_r - E)^2 + \frac{1}{4}\Gamma^2} \exp\left[-\frac{2\pi e^2}{hv}\right], \quad (1)$$

they calculated that the resonance energy would have to be 124.3 keV and the half-width 71.7 keV. However, their values for the cross-section depend critically on the energy-range relation for tritons in heavy ice, about which there is some considerable doubt. To eliminate this they compared the D-T cross-section with that of the $D(d, p)T$ reaction at such an energy that the approaching particles have the same relative velocity in both cases. As their values for the D-T cross-section depend on exactly the same energy-range relation as the values for the D-D cross-section, then the uncertainty should be eliminated. At the same time the Gamow term entering into the D-T cross-section should be, at least to a first approximation, the same as that entering into the D-D cross-section (and which at low energies can be shown to explain completely the variation with energy of the D-D cross-section (Konopinski & Tellar 1948). Thus the ratio between the D-T cross-section and the D-D cross-section should show simply the resonance term in (1) above. On this basis Bretscher, French & Seidl (1948) calculate the resonance to be at 343 keV with a half-width of 80 keV, a result in disagreement with their previous estimate. It was in order to resolve the uncertainty in extrapolating results beyond the experimental points that the present work was undertaken, making use of the higher bombarding energies available from our machine. As the angular distribution of the reaction, as measured by Bretscher & French (1949) has been shown to be symmetrical at energies up to 85 keV triton energy,* it was decided to confine the measurements to observations at 90° to the incident triton beam.

ACCELERATION EQUIPMENT

The high-voltage set follows conventional practice and will not be described in detail. It is a Cockcroft-Walton voltage multiplier giving an output of 200 kV d.c. from an input of 25 kV 1200 c./sec. a.c. The voltage was controlled by a variac feeding the h.t. transformer, and fine control over a limited range was obtained from a rheostat in series with the transformer primary. The maximum peak to peak ripple on the high-tension terminal was measured as 1 to $1\frac{1}{2}\%$ of the d.c. output voltage.

Two methods of measuring the voltage were used. Initially the current flowing down a 2000 M Ω (nominal) resistor was measured with a 0 to 100 μA Unipivot meter. The resistor consisted of a chain of two hundred 10 M Ω , 1 W carbon resistors enclosed in a paxolin tube, and the actual values of the nominal resistors were separately measured. The voltage measured in this way probably carries the greatest error in the experiment.

In view of this, a more reliable method was devised which became available in time for the thin target work. A potential divider was constructed of a chain of two hundred 6.8 M Ω high-stability carbon resistors and connected in place of the resistor previously used. The voltage across the lowest five of these resistors was measured by means of a valve voltmeter. It is our experience that the ratio of such a potential divider varies in use by less than 1%.

A radio-frequency excited ion source similar to those described by Rutherglen & Cole (1947) and Thonemann, Moffat, Roaf & Sanders (1948) was used. The gas

* Subsequent work by the authors indicates that this is also true between 125 and 200 keV.

consumption of this source was 15 ml./hr., and an accelerated beam of $100\ \mu\text{A}$ can be obtained under these conditions. In order to determine the ionic composition of this beam a 15° magnetic analyzer was fitted after the accelerating tube, and analyses of the beam made at frequent intervals throughout the experiment. A typical analysis gives 90 % monatomic, 6 % diatomic and 4 % triatomic ions; the proportion of monatomic ions never fell below 85 % and was rarely below 90 %.

MEASUREMENT OF THE CROSS-SECTION

Preliminary work

A rough determination of the cross-section was made by observing the variation of yield with energy when a thick target of heavy ice was bombarded with tritons. This work (Allan & Poole 1949) showed that the cross-section continues to rise beyond 100 keV incident triton energy, and attains a value of about 5 barns at about 170 keV energy. In view of the considerable uncertainties in the previous experiments caused by the lack of a reliable voltage measurement, and by the necessity to differentiate the experimental curve (thus bringing the voltage critically into the results), it was thought necessary to repeat these measurements using a thin vapour target, instead of the thick-ice target, and using a more reliable method to measure the voltage.

Method

To perform a thin target measurement the T beam was passed via a system of traps and canals, into a chamber containing D_2O vapour at a few tenths of a mm. pressure, and the alpha-particles emitted from a defined portion of the track were observed in a direction at right angles to the beam with a proportional counter. The presence of the vapour in the target chamber makes a measurement of the beam current very difficult, and the following scheme, using the D-D reaction as a monitor, was adopted to avoid the need for such a measurement.

A quantity of deuterium was mixed with the gas fed to the ion source, and the protons from the D-D reaction were observed simultaneously with the alpha-particles from the T-D reaction, using the same counter.

Then, if n_1 = no. of T atoms/sec. in the beam,

n_2 = no. of D atoms/sec. in the beam,

σ_1 = T-D cross-section,

σ_2 = D-D (d, p) cross-section,

ν_1 = no. of T-D disintegrations/sec.,

ν_2 = no. of D-D disintegrations/sec.,

$$\frac{\nu_1}{\nu_2} = \frac{n_1 \sigma_1}{n_2 \sigma_2}.$$

As the observations are always made at 90° it is only the isotropic portion of the D-D cross-section which is of importance here. The data required for this are taken from the summarized data on the D-D reaction contained in Konspinski & Teller (1948).*

* The data published by Bretscher *et al.* have been recalculated by French to take account of an error in the energy-range originally used. It was these corrected values, communicated privately, that were used.

In order to determine n_1/n_2 , use may be made of the thick target yields for the D-D and D-T reactions given by Bretscher, French & Seidl (1948) and Bretscher & French (1949) by bombarding a thick heavy-ice target with the mixed beam at some definite voltage (110 keV was always used) and observing the ratio between the yields of alpha-particles and protons. The error involved in this use of a thick target is quite small, as no differentiation or energy-range relation is involved.

To perform the experiments, hydrogen containing about 0.2 % of tritium and about 30 % of deuterium was admitted to the ion source and the resulting beam allowed to enter the reaction chamber without magnetic analysis. Although resolution of the beam would have been desirable, it was not feasible for the following reasons. The beam will contain ions of the compositions: H^+ ; H_2^+ , D^+ ; H_3^+ , $(\text{HD})^+$, T^+ ; $(\text{H}_2\text{D})^+$, $(\text{HT})^+$, D_2^+ ; $(\text{H}_2\text{T})^+$, $(\text{HD}_2)^+$, $(\text{DT})^+$, ..., etc. Of these the mass 4 beam is the ideal one to use, thereby enabling one to compare the yields from the D_2^+ and the $(\text{HT})^+$ components, with interference only from the $(\text{H}_2\text{D})^+$ component, which may be expected to be weak. The $(\text{D}_2)^+$ and $(\text{HT})^+$ components being both diatomic ions may be expected to arise in a similar manner in the ion source and so to maintain a constant ratio in intensity. However, if this beam is used the effective energy of the triton is only three-quarters of the energy available which would be a severe limitation on the range of the experiment. The mass 3 beam is the obvious alternative, but this implies comparing the yield from the atomic T^+ beam with that from the molecular $(\text{HD})^+$ beam, and experience shows that the ratio between the monatomic and diatomic components depends markedly on the running condition of the source. Therefore the total beam was used, and it was assumed that the effect of the T^+ and D^+ components would swamp that of all the others because of the high percentage of atomic ions present. How true this is is discussed in the section on errors below.

Apparatus

A diagram of the target chamber used for the experiment is shown in figure 1. The beam enters from the left, passing first through diaphragm *A* (3mm. in diameter) and continuing down tube *B* to enter the reaction space *K* through the diaphragm system *C*. Space *K* is filled to a pressure of a few tenths of a millimetre of mercury of D_2O vapour, admitted via tube *J*. The vapour passes out through *C* and is finally trapped by tube *B*. This system was found in practice to be very efficient, no noticeable amount of vapour passing back into the acceleration tube.

Disintegration particles (protons and alpha-particles) are allowed to reach the counter through a rectangular diaphragm *D* which lets the counter *E* 'see' a 1 cm. length of the beam, and serves to limit the angular divergence of particles entering the counter. A shutter *F* could be turned to obscure the window *D* to allow observations of pulses produced in the counter by stray neutrons or gamma rays.

A liquid-air container mounted on a Wilson seal was mounted above the counter, and carried on its lower side a copper rod on which a thick heavy-ice target could be deposited, and by rotation about the Wilson seal, swung into the line of the beam. This was used to determine the value of n_1/n_2 as previously mentioned. When not in use the target rod was moved to a position behind the counter.

A proportional counter was used to record the disintegration particles. This was constructed of brass and was 2 cm. in diameter, with a 0.2 mm. diameter tungsten wire. Fillings of argon with 5 to 10 % of carbon dioxide were used, to a total pressure of 23 cm. (and later 11.5 cm.). Discriminator bias curves obtained (at either pressure) by testing with a plutonium alpha-particle source placed 2 cm. from the windows exhibited a perfectly flat plateau. Two counters were made and exhibited substantially identical characteristics. Three 3 mm. diameter mica windows were

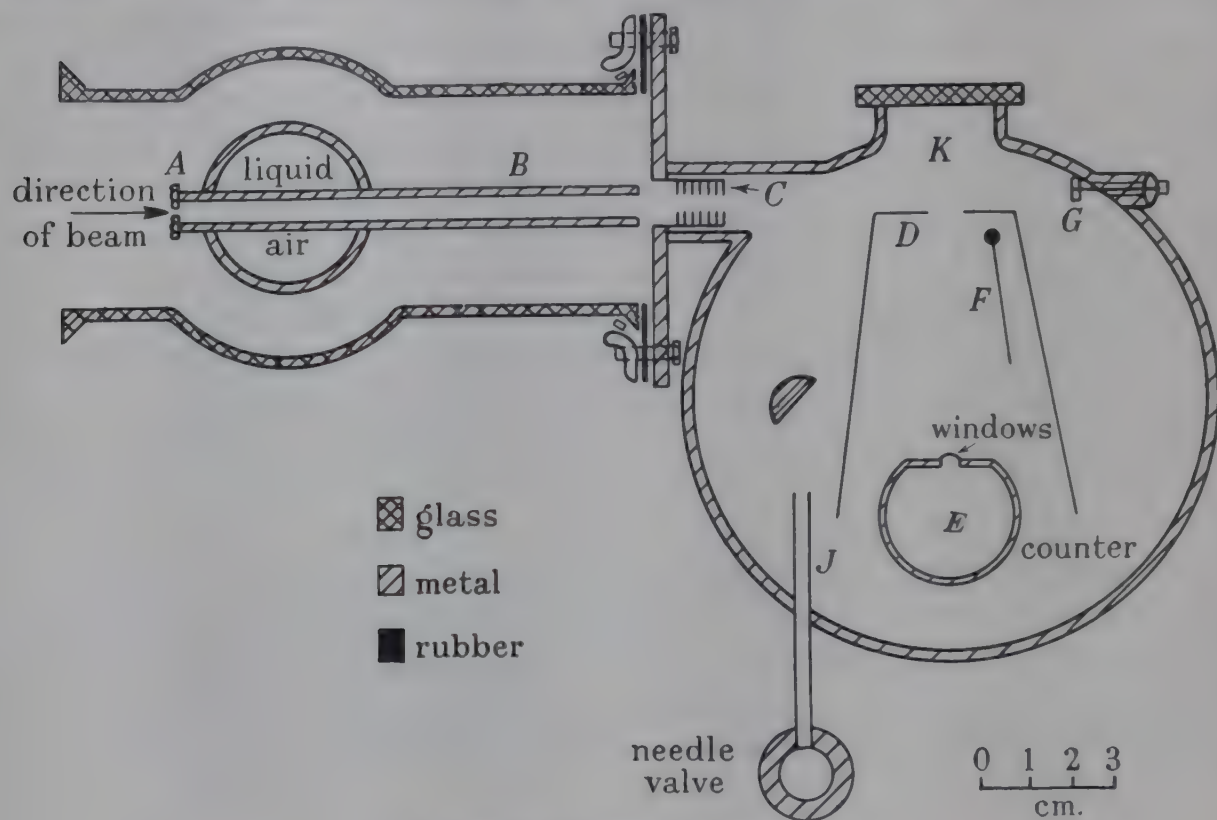


FIGURE 1. The target chamber.

provided, the mica being affixed with Apiezon wax W and a brass support plate with corresponding holes clamped over the window to enable it to withstand a pressure difference in either direction. The thickness of the mica window was carefully chosen, since it was desired to pass the alpha-particles from the T-D reaction (energy 3.5 MeV, range 2.1 cm.) and the protons from the D-D reaction (energy 3 MeV, range 14 cm.) but not the tritons from the D-D reaction (energy 1 MeV, range uncertain but about 1.3 cm.). A window of 1.7 cm. air equivalent was used, its thickness being determined by measuring the residual range of polonium alpha-particles after passing through the mica.

This counter was connected through a single-valve head amplifier to a linear amplifier. The amplified pulses could then be fed into either or both of two discriminators, or a four-channel pulse analyzer. The pulse analyzer was used initially to plot the bias curve for protons, in order to save the time involved in using a discriminator. The gain of the amplifier was set so that the peak of the pulse-height distribution arising from the protons occurred at about 11 V. Under these conditions if one discriminator was set at a low voltage (usually of the order of 7.5 V) it counted all of the protons and alpha-particles but very little of the background due to

neutron recoils (or gamma-rays), while if the other was set to 30 V it counted all of the alpha-particles but not the protons. From these two readings the proton count alone could be deduced by subtraction. As the two counts were made simultaneously on pulses from the same counter this procedure did not lead to a poor statistical accuracy. During most of the runs the pulse analyzer was also connected to plot the distribution of pulses from the protons in order to check that there had been no shift in the position of the peak.

Performance of the apparatus

In order to determine the spread of pulse heights arising from the protons, the following experiments were carried out. A polonium alpha-particle source was secured to the opening in *D* (figure 1), and the amplifier gain and counter voltage adjusted until the mean pulse height was about 11 V (as for the proton pulses). Pulse analyzer curves of these pulses were done with slightly varying counter voltages, a typical one being shown in figure 2*a*. These are taken as the basis for assessing the curves obtained from the D-D protons.

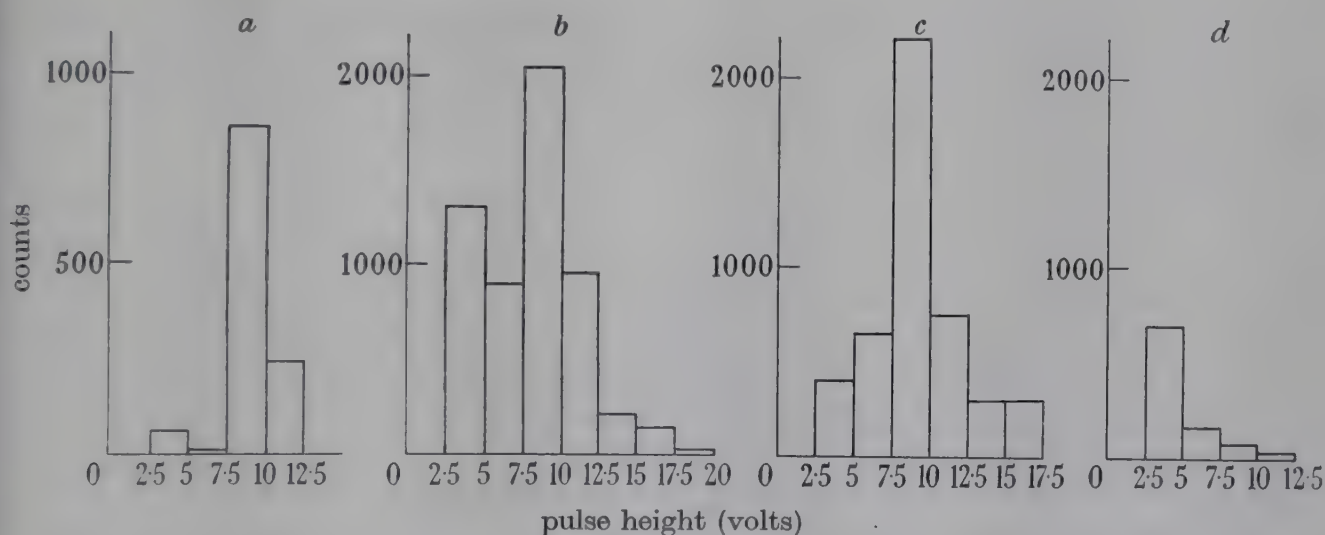


FIGURE 2. Pulse-height distributions. *a*, polonium alpha-particles; *b*, *c*, protons; *d*, background.

The first runs were carried out without the diaphragm *A* (figure 1) to collimate the beam initially, so that a considerable proportion of the initial beam struck the first diaphragm in *C*. The pulse analyzer curves of the proton pulses taken in this way (of which a typical example is shown in figure 2*b*) showed a considerable background of small pulses. These were ascribed to recoils in the counter caused by neutrons generated at *C* by the beam stopped there. In order to reduce this effect the diaphragm at *A*, 3 mm. in diameter, was introduced, thus considerably reducing the neutron flux at the counter. The curve shown in figure 2*c* is a pulse analyzer curve of the proton pulses taken after fitting the diaphragm at *A*. It will be seen that although the background is reduced, it is not yet entirely eliminated. In order to determine the spectrum of the pulses caused by neutrons the shutter *F* was closed and a pulse-height distribution of the residual pulses plotted. This is shown at figure 2*d* (the vertical scales are not directly comparable as the different pulse analyzer runs required to build up the whole curve were simply monitored with a discriminator

counting all pulses above a fixed bias). From this it is immediately obvious that even after allowing for the effect of neutron produced recoils, the pulse-height distribution for protons is very considerably broader than that for the polonium alpha-particles. This phenomenon has not been satisfactorily accounted for. However, these curves enabled us to decide on a discriminator bias at which substantially all of the protons could be counted. Simple inspection of an oscilloscope screen showed that the pulses from the alpha-particles saturated the amplifier at an amplitude of about 60 V. Accordingly, the second discriminator was set at 35 V bias.

Counting procedure

Several runs were made, the procedure in all cases being similar, as follows:

(1) The approximate amount of deuterium to mix with the tritium in order to get nearly equal counting rates from the two reactions was calculated. This deuterium was sealed in a glass bulb, attached to the palladium leak of the high-tension set along with bulb containing the dilute tritium, and the gases caused to mix by breaking fragile glass seals.

(2) The set was turned on and a heavy-ice target deposited in the target chamber. The target was turned into the path of the beam, and with the set running at 110 kV the numbers of alpha-particles and of protons were counted. Counts were continued during the arrival of 10^4 alpha-particle counts.

(3) Without turning off the set the heavy-ice target was removed by warming up the liquid-air trap, and its carrier rotated out of the beam. The needle valve controlling the flow of heavy-water vapour was then opened to its fullest extent. This established a pressure of a few tenths of a millimetre of mercury in the target chamber. Counts were now taken of the protons and alpha-particles arriving in the counter for various accelerating voltages. Usually counting was continued during the arrival of 2000 alpha-particle pulses, and usually two counts were made at each voltage.

(4) Without turning off the set a new heavy-ice target was deposited, and the thick target ratio between the alpha-particles and protons once again determined at 110 kV.

The time taken for each run was approximately $1\frac{1}{4}$ hr.

Results

The experiment described gives the ratio between the cross-sections for the T-D reaction and the $D(d, p)$ reaction at equal bombarding energies. In order to obtain the T-D cross-section this ratio is multiplied by the published values for the D-D cross-section (Bretscher *et al.* 1948; Konopinski & Teller 1948; see also footnote, p. 490). The results obtained in this way are shown in figure 3, some of the low-energy values obtained by Bretscher & French (1949) being shown on the same graph. It will be seen that the junction of our results with those of Bretscher & French is smooth, and the results are not inconsistent with those of Baker *et al.* (1948).

Errors

The accuracy of these experiments will be affected in the following ways:

(a) Statistical accuracy of counting. The errors arising from this cause are $\pm 2\%$.

(b) The percentage of tritium ions in the beam varied during the experiment, as shown by the thick target counts made before and after each run. The magnitude of this variation was always about 5 %, and always represented a decrease in the amount of tritium present. No correction has been made for this variation, but as the various points in each run were taken in different orders (voltage increasing, voltage decreasing and alternately high and low voltages) and subsequently all values were averaged, the error introduced will tend to cancel and so will be less than 5 %.

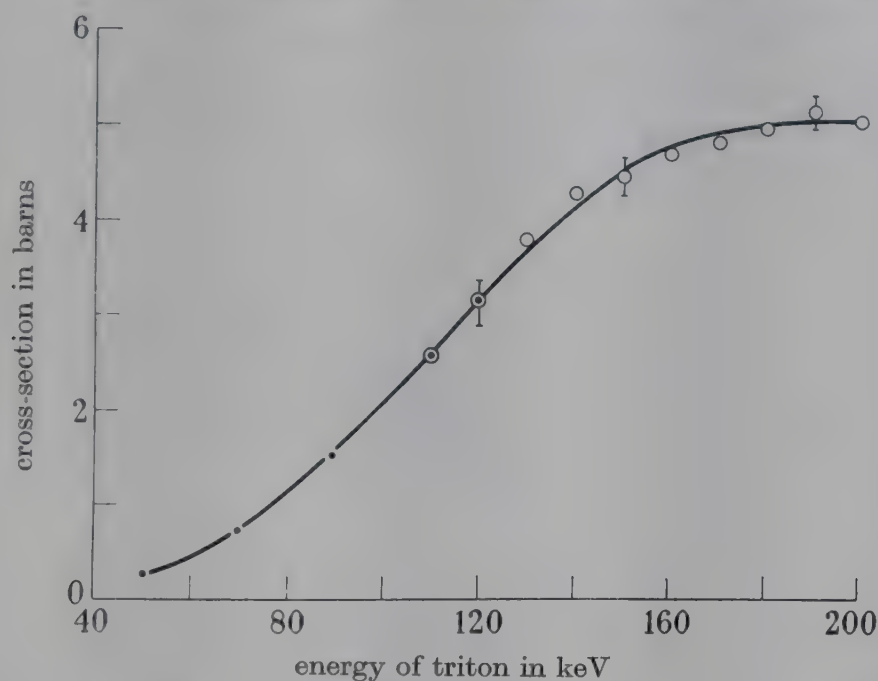


FIGURE 3. The T-D cross-section. • results of Bretscher & French (1949);
○ results of present work.

(c) A small percentage of 'background' counts was still included in the proton group. It is thought that these will not amount to more than 2 % of the whole.

(d) The values for the cross-section depend on the value of the D-D cross-section as published, which in its turn depends on the energy-range relation for deuterons in heavy ice. While there is a possibility of an error up to ± 5 % arising from this source, it seems unlikely that there is any greater uncertainty, particularly in view of the close agreement at low energies between the values of Bretscher *et al.* and the thin target result of Sanders, Moffatt & Roaf (1950).

Thus the intrinsic errors of the experiment, represented by (a), (b) and (c) above, may amount to ± 8 %. In addition, there may be a systematic error in the values of the D-D cross-section used. It is difficult to estimate the amount of this, but it is unlikely that it will amount to more than 15 %.

Discussion of the results

In discussing these results the following considerations must be borne in mind.

(1) As all the measurements were made at 90° to the beam, then the cross-sections given represent 4π times the differential cross-section at 90° . If the angular distribution is completely isotropic, then this is the total cross-section for the reaction: if not, the value given may be either smaller or larger than the total cross-section.

The cross-section as measured will be referred to as the 'isotropic' cross-section for the reaction.

(2) The observations were all made at 90° in the laboratory co-ordinate system. The correction to be applied to convert this to a system in which the centre of gravity of the particles is at rest is never greater than 2 %. It has therefore been neglected.

(3) These measurements were made with the total beam, which in fact contained a molecular component between 10 and 15 %. In order to assess the error introduced in this way a cross-section curve somewhat similar to our experimentally observed one was assumed, and from this the values that would be observed were calculated for the case when a 10 % molecular component was present, and for the case when 15 % was present. These curves are plotted in figure 4, and it will at once be seen that the error from this cause is not serious.

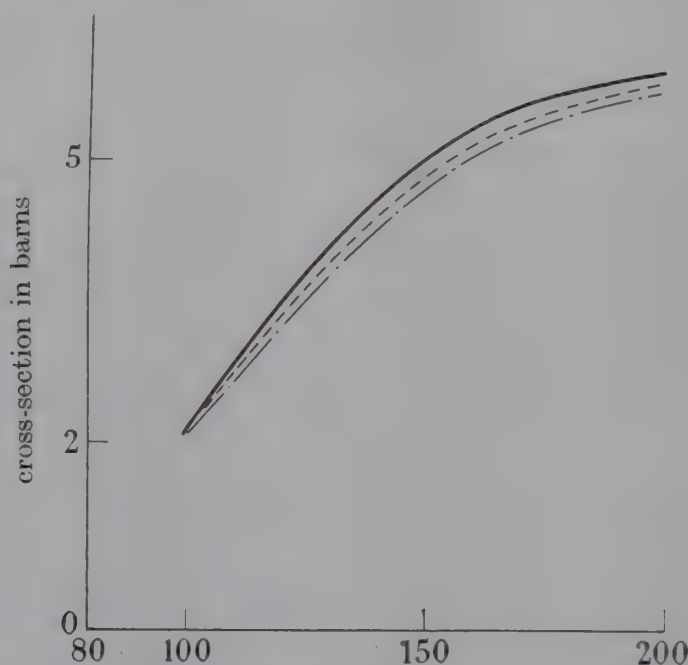


FIGURE 4. Errors due to molecules in the beam. — assumed value; --- effect of 10 % molecules; —. — effect of 15 % molecules.

Konopinski & Teller (1948) considering the D-D reaction, have shown that the cross-section may be made up of terms of the form

$$\sigma_l = \pi \lambda^2 (2l + 1) g_l |\alpha_l|^2 P_l,$$

where σ_l represents that part of the cross-section due to pairs of particles approaching with orbital number l , g_l is a statistical term depending on spins, P_l is the probability of the two incident particles approaching close enough to interact (the penetrability), and $|\alpha_l|^2$ is the 'intrinsic reaction probability' which includes any resonance term; both $|\alpha_l|^2$ and P_l must be positive and less than 1.

Both Bretscher & French (1949) and Taschek *et al.* (1948) have assumed that only incoming S -waves are effective (presumably on the grounds of the isotropic angular distribution) and have put $|\alpha_l|^2$ into the form of a Breit-Wigner resonance and P_l as the low-energy approximation given by Bethe (1937). In this way they obtained formulae of the form

$$\sigma(E) = \frac{A}{E} \frac{\exp\left[-\frac{2\pi e^2}{\hbar v}\right]}{(E_r - E)^2 + \frac{1}{4}\Gamma^2}.$$

In figure 5 this function is plotted, putting in the constants according to Bretscher and according to Taschek, and, in addition, a formula of this type computed by the authors, but using a different value for E_r and Γ , and using the more exact S -wave

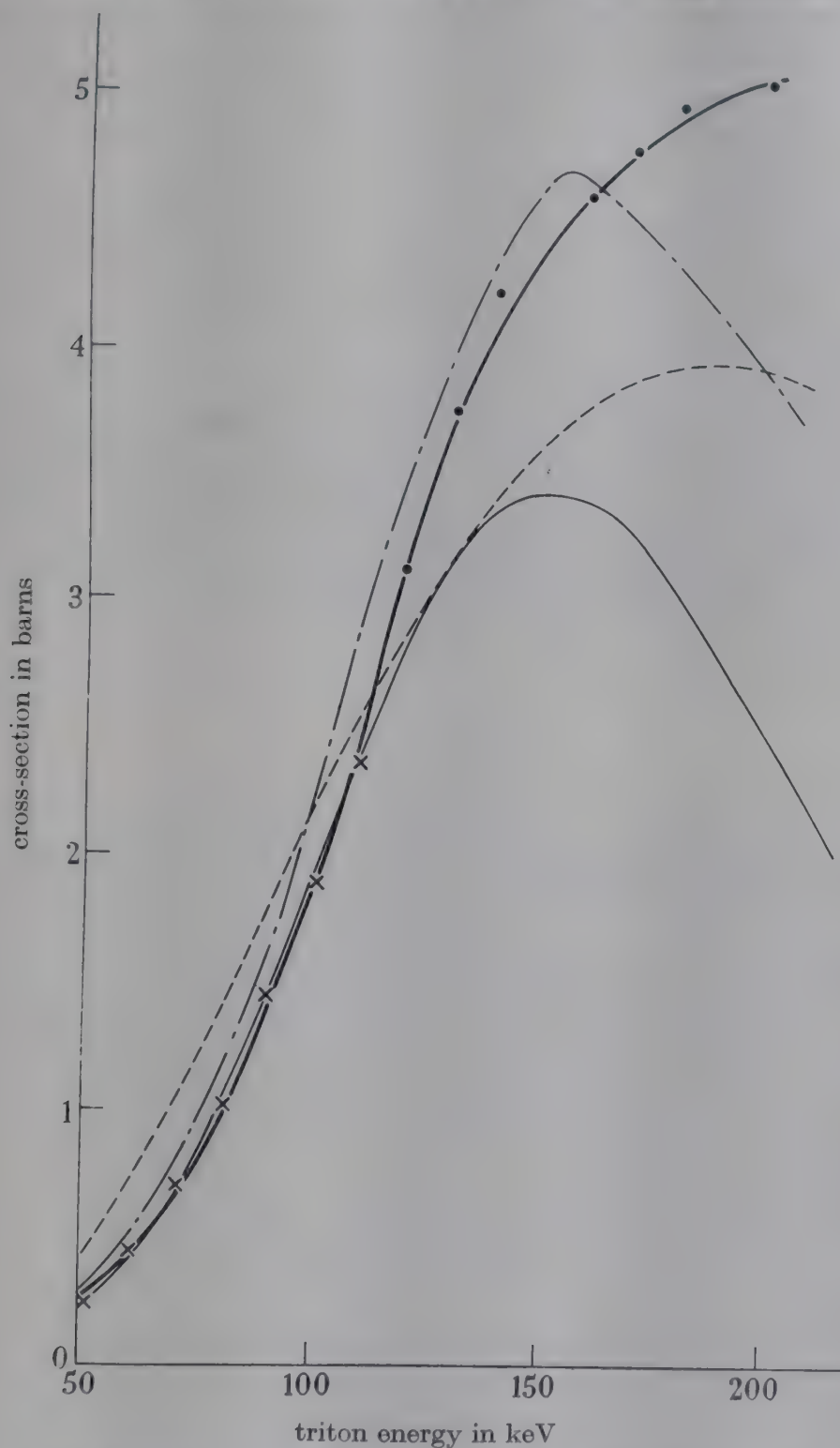


FIGURE 5. T-D cross-section. —•— authors' points; $\times$ Bretscher's points; — Bretscher's resonance formula; --- Taschek's resonance formula; -.- authors' resonance formula.

penetrability as given by Bethe (1937, equation (8)). It can be seen that none of these curves fits the data well, particularly in the high-energy region. However, a far more serious objection to this kind of treatment arises when the absolute magnitude of the cross-section is considered. At 200 keV, $\pi\lambda^2 = 6.1$ barns and the S -wave pene-

trability = 0.35. Thus, according to Konopinski & Teller's formula above the maximum possible cross-section due to S -waves is 2.1 barns. The observed cross-section is 5.14 barns. This matter will now be considered in a little more detail.

The incoming system consists of pairs of dissimilar particles of spin $\frac{1}{2}$ and 1 respectively. Doublet and quartet states may therefore be formed which may combine with any value of orbital momentum.* The outgoing particles have spin $\frac{1}{2}$ and 0 and so can form only doublet states. Thus in the absence of spin-orbit coupling, and neglecting all values of the initial orbital quantum number greater than 1, the only transitions that are allowed by the conservation laws of angular momentum and parity are ${}^2S \rightarrow {}^2S$ and ${}^2P \rightarrow {}^2P$ (the triton, deuteron and alpha-particle are all taken to have even intrinsic parity in their ground states). Of these ${}^2P \rightarrow {}^2P$ will not contribute to the isotropic cross-section as measured here. If spin-orbit coupling is allowed then ${}^4P \rightarrow {}^2P$ and ${}^4S \rightarrow {}^2D$ may also take place, both of which will contribute to the isotropic cross-section, and, in addition, it will now be possible for the component ${}^2P_{\frac{1}{2}} \rightarrow {}^2P_{\frac{1}{2}}$ of the ${}^2P \rightarrow {}^2P$ transitions to affect the isotropic cross-section. (Effectively the presence of spin-orbit coupling here means that the intensities of the various components of the ${}^2P \rightarrow {}^2P$ transitions are no longer necessarily in purely statistical proportions, so the angular distribution resulting is no longer $\cos^2 \theta$.)

If only the transitions ${}^2S \rightarrow {}^2S$ are taken into account, then the cross-section becomes

$$\sigma = \frac{1}{3} \pi \lambda^2 P_0 |\alpha_0|^2,$$

where P_0 and $|\alpha_0|^2$ are each less than one. At 200 keV triton energy the value of $\pi \lambda^2$ is 6.1 barns. Thus it is at once obvious that the value of the cross-section at 200 keV cannot possibly be explained by taking account of 2S transitions only.

If now the transitions ${}^4S \rightarrow {}^2D$ are included it becomes possible to achieve the observed cross-section, but, as pointed out above, a further difficulty arises when the value of the penetrability is calculated. If it is assumed, as for the D-D reaction, that reaction takes place at a distance of 7×10^{-13} cm., then the S -wave penetrability at 200 keV is 0.35, while the P wave penetrability is 0.02. Thus the total available cross-section from 2S and 4S transitions is 2 barns, and that from ${}^4P \rightarrow {}^2P$ and ${}^2P_{\frac{1}{2}} \rightarrow {}^2P_{\frac{1}{2}}$ cannot exceed 0.4 barn.

From this it is clear that unless some new mechanism is postulated whereby the reaction can take place without penetration of the potential barrier, either the use of the penetrability calculated as a W.K.B. approximation must be abandoned,† or the reaction must take place when the incident nuclei are much further apart than 7×10^{-13} cm. This would correspond to a large radius for the ${}^5\text{He}$ compound nucleus, probably of the order of 15×10^{-13} cm. Until this difficulty is solved there seems little point in considering whether or not a resonance is required to produce the observed curve. In addition it seems clear that no mechanism can explain the large cross-

* There is some doubt as to whether S -states can be formed at all if this involves close approach of all the particles, as this would seem to violate the Pauli principle. See Konopinski & Teller concerning the quintet state in the D-D reaction.

† Mr B. H. Flowers of this establishment has made calculations on this basis which give better agreement with the experiments.

section without assuming strong spin-orbit coupling to enable quartet transitions to become effective. The position may then be summarized as

(1) No formula taking account of incident S -waves only can account for the high value of this cross-section if the use of penetrability theory as given by Bethe is retained. However, there has always been considerable doubt as to whether this theory is applicable to such light nuclei.

(2) The high value of the cross-section provides evidence for spin-orbit coupling in this reaction.

(3) The evidence for a resonance is inconclusive.

(4) It is a matter of importance to know the angular distribution between 85 keV (at which it is isotropic) and 200 keV (see footnote, p. 489).

The authors wish to acknowledge their indebtedness to the following: (1) The Director, Atomic Energy Research Establishment, Harwell, for permission to publish this work. (2) Dr E. Bretscher, for his support and encouragement in this work. (3) Dr W. Arrol, who extracted and purified the tritium from the lithium fluoride. (4) The Isotopes Division, A.E.R.E., who irradiated the lithium fluoride required for the thick target experiment. (5) National Research Council of Canada, for providing the irradiated lithium fluoride required for the thin target experiment. (6) Professor M. H. L. Pryce, for much helpful criticism of the theoretical interpretation.

REFERENCES

- Allan, D. L. & Poole, M. J. 1949 *Nature*, **164**, 102.
 Baker, C. P., Holloway, M. G., King, L. D. P. & Schreiber, R. E. 1948 Declassified Report A.E.C.D. no. 2226.
 Bethe, H. A. 1937 *Rev. Mod. Phys.* **9**, 178.
 Bretscher, E. & French, A. P. 1949 *Phys. Rev.* **75**, 1154.
 Bretscher, E., French, A. P. & Seidl, F. G. P. 1948 *Phys. Rev.* **73**, 815.
 Konopinski, E. J. & Teller, E. 1948 *Phys. Rev.* **73**, 822.
 Rutherglen, J. G. & Cole, J. F. I. 1947 *Nature*, **160**, 545.
 Sanders, J. H., Moffatt, J. & Roaf, D. 1950 *Phys. Rev.* **77**, 754.
 Taschek, R., Everhart, G., Gittings, H. A., Hemmendinger, A. & Jarvis, G. 1948 Declassified report A.E.C.D. no. 2250.
 Thonemann, P. C., Moffat, J., Roaf, D. & Sanders, J. H. 1948 *Proc. Phys. Soc.* **61**, 15.

Experiments on the reaction $T(d, n)^4\text{He}$

II. The angular distribution

BY D. L. ALLAN AND M. J. POOLE

Atomic Energy Research Establishment, Harwell, Berks

(Communicated by Sir John Cockcroft, F.R.S.—Received 19 July 1950—

Revised 19 August 1950)

A measurement of the angular distribution of α -particles from the T-D reaction has been made by bombarding a heavy-water vapour target with a beam of tritons. Observations were made at three triton energies, 125, 175 and 200 keV, but only with high statistical accuracy at 175 keV. In all cases the disintegration was found to take place symmetrically in centre of mass space within the limits of experimental error.

INTRODUCTION

In view of the extremely high cross-section already reported for the T-D reaction (Bretscher & French 1949; Allan & Poole 1951), it was thought of interest to continue the measurements by observing the angular distribution of the emitted α -particles with respect to the direction of emission of the beam. Bretscher & French (1949) have already measured this at 35 keV and with more accuracy at 75 keV, while Taschek, Everhart, Gittings, Hemmendinger & Jarvis (1948) made measurements between 1 and 2.5 MeV and found the cross-section could be represented by a formula of the form

$$\sigma(\theta) = a + b \cos \theta + c \cos^2 \theta,$$

where a , b and c are energy dependent parameters.

APPARATUS

To perform this experiment the apparatus shown in figure 1 was constructed and attached to the 200 kV high-tension set. A beam of tritons, separated from molecules by the magnetic analyzer, entered through tube A which is at liquid-air temperature and serves to condense the heavy water issuing from the target region. Initial collimation is performed at D and prevents any beam from striking the sides of B and setting up a neutron source there. Once inside the chamber the beam passes along tubes E and F and finally hits collector G . Tubes E and F have their facing edges turned down to a fine knife edge, and are mounted so that their axes lie accurately along a diameter of the chamber with the gap between them at its centre.

Alpha-particles originating between the ends of E and F were observed in two counters. One of these was fixed to observe particles at 90° to the beam direction, while the other was on an arm mounted in ball races so that it could be moved to observe particles emitted at any angle between 30° and 150° . This arm entered the chamber through a Wilson seal and was hollow, the lead to the counter wire being run inside it. The counters were side-window proportional counters with twelve

3.5 mm. diameter windows arranged in a honeycomb pattern. The window of the movable counter subtended about 5° at the target, which therefore represents the best angular resolution obtainable. Mica of 1.0 cm. air equivalent or less was used because the α -energy progressively diminishes as the angle of observation is increased,

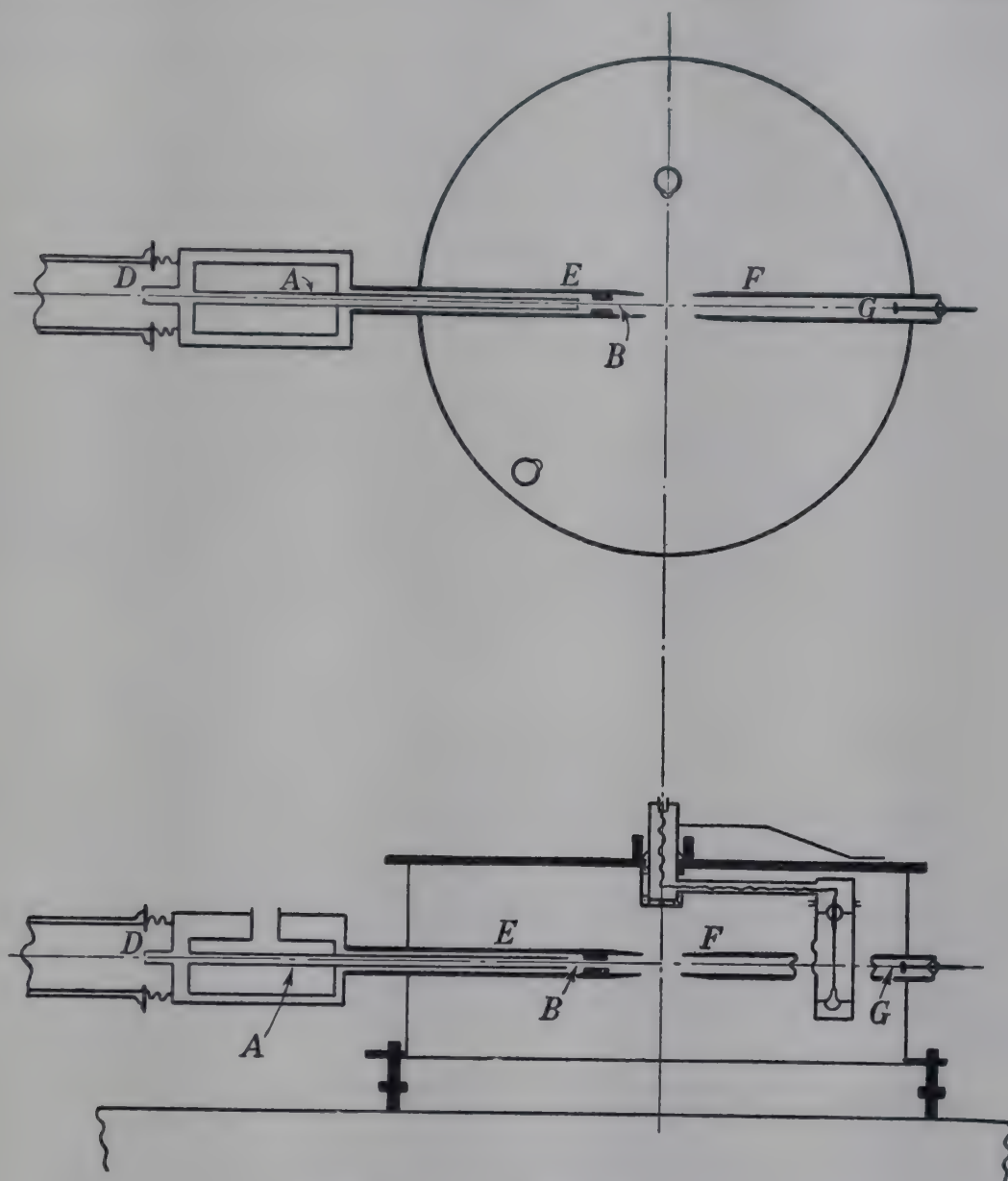


FIGURE 1. Angular distribution tank.

falling to 2.5 MeV for observation at 150° with 200 keV bombarding energy. This mica was carefully measured before fitting to the counter, and further checks were made with an α -particle source after filling, and finally, by noticing the highest angle at which tritons from the D-D reaction could be observed.

All measurements were taken as ratios between the count in the movable counter and that in the fixed counter. Therefore it was possible to use the mass 3 beam, containing triatomic hydrogen ions as well as tritons, as the actual value of the beam current is not significant as long as the 'foreign' ions do not produce disintegration particles. To save time during the runs the magnetic field required for the mass 3 beam was first determined by running the ion source on a mixture of hydrogen and deuterium and setting up on the $(\text{HD})^+$ beam.

EXPERIMENTAL PROCEDURE

The following routine was used in all runs.

(a) The set was turned on using a mixture of hydrogen and deuterium in the ion source and the magnetic field required to cause the $(\text{HD})^+$ beam to enter the tank was determined.

(b) The set was changed over to a tritium-hydrogen mixture. Since two palladium leaks were fitted this simply involved turning a switch.

(c) A series of readings was taken round the target. Where only a rough run was required, readings were taken every 30° . For a more accurate run readings were taken every 20° or 15° .

RESULTS

A number of runs were made in this way at 175 keV bombarding energy, and single runs at 125 and 200 keV. These results, corrected for the transformation from laboratory to centre of mass space and for a small variation of the target volume with angle, are shown in figure 2. It will be seen that at all these energies the reaction is isotropic within the accuracy of the experiment. These experiments together with those of Bretscher & French therefore establish approximate isotropy from 35 keV bombarding energy to 200 keV energy.

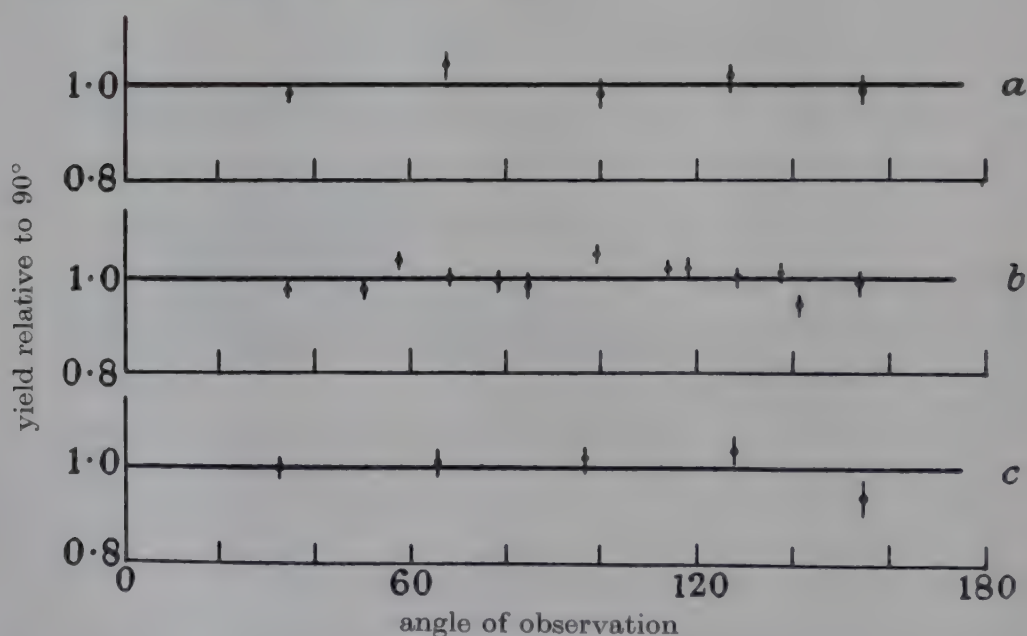


FIGURE 2. Angular distributions of α -particles from the reaction $\text{T}(d, n)^4\text{He}$ at triton energies: *a*, 200 keV; *b*, 175 keV, *c*, 125 keV.

The authors must acknowledge their indebtedness to the Director, Atomic Energy Research Establishment for permission to publish this work.

REFERENCES

- Allan, D. L. & Poole, M. J. 1951 *Proc. Roy. Soc. A*, **204**, 488.
 Bretscher, E. & French, A. P. 1949 *Phys. Rev.* **75**, 1154.
 Taschek, R., Everhart, G., Gittings, H. T., Hemmendinger, A. & Jarvis, G. 1948 Declassified report *A.E.C.D.* no. 2250.

The theory of the T+D reaction

By B. H. FLOWERS, *Atomic Energy Research Establishment, Harwell*

(Communicated by Sir John Cockcroft, F.R.S.—Received 4 August 1950)

The energy dependence and absolute magnitude of the cross-section for the T+D reaction are described in terms of the exact wave-functions of the Coulomb field and a complex reaction length defined by the logarithmic derivative of the wave-function on the reaction surface. Previous difficulties in the interpretation of the experiments are largely attributed to the failure of the Wentzel-Kramers-Brillouin approximation. It is sufficient, but not necessary, to assume that the whole cross-section is due to *s*-waves; but in either case the presence of strong spin-orbit coupling is indicated.

1. INTRODUCTION

It is customary to analyze yield curves for nuclear reactions involving charged particles in terms of a Gamow penetration factor G , representing the repulsion of the electrostatic field, together with some intrinsic reaction probability $Q(E)$ which may be, for example, a simple resonance formula. In terms of these quantities the reaction cross-section may be written as a sum of partial cross-sections σ_l for incident waves of angular momentum l in the form

$$\sigma = \sum_l (2l+1) \sigma_l,$$

where

$$\sigma_l = \omega_l \pi \lambda^2 G_l Q_l(E). \quad (1)$$

Here $2\pi\lambda$ is the de Broglie wave-length and ω is a statistical weighting factor dependent upon the spins of the initial particles and the relative angular momentum. Both G and Q are normally restricted to be less than unity. This, however, is an unnecessary restriction, since it is only the product GQ which may never exceed unity; moreover, only in rather special circumstances is it possible to separate the two quantities.

The penetration factor G is calculated with the aid of the Wentzel-Kramers-Brillouin (WKB) approximation (Bethe 1937), the usefulness of which in problems of alpha-radioactivity and many nuclear reactions has been well demonstrated. Its validity depends upon the assumption that the closest distance of approach of two particles in the Coulomb field is large compared to the natural electrostatic unit of length of the system, the effective Bohr radius, given by $d = \hbar^2/2z_1z_2e^2\mu$, where z_1e and z_2e are the charges and μ the reduced mass of the system. For problems involving heavy nuclei, owing to the charge dependence of d , this natural unit of length is small, of the order of 10^{-13} to 10^{-14} cm., whereas the closest distance of approach is at least the nuclear radius, of the order of $(5 \text{ to } 10) \times 10^{-13}$ cm. Under these circumstances the WKB method gives a fair approximation to the real wave-functions. In reactions involving only the lightest nuclei, however, the position may be reversed: thus, in the reaction of immediate concern



we have $d = 12.2 \times 10^{-13}$ cm., while the reaction radius, defined to be the maximum separation at which nuclear forces can operate and therefore give rise to a reaction, is about 7×10^{-13} cm., compounded from the radii of deuteron and triton plus the range of nuclear forces. (The precise definition of this quantity will be discussed further at a later stage.)

Thus the WKB method might be expected to give a poor approximation to the true wave-function in the region of the reaction radius, and therefore a poor approximation to G in this reaction. (The detailed numerical comparison of the exact and the WKB wave-functions has been performed by Yost, Wheeler & Breit (1936)).

The intrinsic reaction probability is often assumed to be in the form of a resonance:

$$Q(E) = \frac{\text{const.}}{(E - E_0)^2 + \frac{1}{4}\Gamma^2}, \quad (3)$$

where E is the energy of the incident particle, E_0 its energy at resonance, and Γ is a constant, the width of the resonance. Only if Γ is very small compared with E_0 can this factor be separated from the penetrability G . If the resonance is a broad one it is necessary to take into account the variation of Γ across the resonance, since Γ itself then involves the penetrability. It is, in fact, rather difficult to define in conventional terms precisely what is meant by a broad resonance, at least for charged particles. In general, we conclude that it is not permissible to separate the quantities G and Q .

The experimental cross-section for the T + D reaction (Bretscher & French 1949; Allan & Poole 1949, 1951) is an example of this. It reaches a maximum value at about 200 keV bombarding triton energy (80 keV in the centre of gravity system), this maximum value being very close to $\pi\lambda^2$, the maximum possible cross-section for s -waves; and if interpreted as a conventional resonance the width has to be taken as about 170 keV in order to get the roughest agreement with the resonance formula. We may therefore expect that it is not permissible to separate the quantities G and Q for this reaction.

The present paper, following a suggestion by Professor M. H. L. Pryce, attempts to calculate the product GQ treated as a single entity. It replaces the WKB approximation by the exact wave-functions of the Coulomb field, and in place of the resonance concept is introduced a complex reaction length analogous to the scattering length of Fermi & Marshall (1947), in terms of which the variation of cross-section with energy over the whole measured region can be established. In the present state of knowledge it is impossible to predict with any certainty the magnitude of this reaction length; its real and imaginary parts therefore form the two unknown parameters of the theory which have to be found by comparison with observation.

2. CHARACTER OF THE TRANSITIONS

A deuteron of spin 1 and a triton of spin $\frac{1}{2}$ may collide in doublet or quartet states. Thus the initial states of the system are

$$^2S, \quad ^4S, \quad ^2P, \quad ^4P, \quad ^2D, \quad \dots$$

The final states are to be compounded from an alpha-particle of zero spin and a neutron of spin $\frac{1}{2}$. They are therefore all doublet states

$$^2S, \quad ^2P, \quad ^2D, \quad \dots$$

In the absence of spin-orbit coupling the following are the only possible transitions:

$$^2S \rightarrow ^2S, \quad ^2P \rightarrow ^2P, \quad ^2D \rightarrow ^2D, \dots,$$

spin and orbital momentum being independently conserved. If spin-orbit coupling is permitted so that spin changes may occur, the additional transitions allowed by conservation of parity and total angular momentum are

$$^4S \rightarrow ^2D_{\frac{1}{2}}, \quad ^4P_{\frac{1}{2}} \rightarrow ^2P_{\frac{1}{2}}, \quad ^4P_{\frac{3}{2}} \rightarrow ^2P_{\frac{3}{2}}, \quad ^4P_{\frac{3}{2}} \rightarrow ^2F_{\frac{5}{2}}, \dots,$$

which might be thought *a priori* to be weaker than the former ones.

The angular distribution of the emitted particles has been measured at 34, 83, 125, 175 and 200 keV by Bretscher & French and by Allan & Poole and has been found to be isotropic at these energies. Accidental isotropy is unlikely to be independent of bombarding energy, and consequently one is led to admit only those transitions which would lead to isotropic emission, namely,

$$^2S \rightarrow ^2S, \quad ^4S \rightarrow ^2D_{\frac{1}{2}}, \quad ^2P_{\frac{1}{2}} \rightarrow ^2P_{\frac{1}{2}}, \quad ^4P_{\frac{1}{2}} \rightarrow ^2P_{\frac{1}{2}}, \quad ^4D_{\frac{1}{2}} \rightarrow ^2S.$$

(Here we have made use of the fact that transitions in which the total angular momentum J is $\frac{1}{2}$ are always isotropic.) Of these transitions the first is expected to be predominant, since it is the only one not requiring spin-orbit coupling, while the last three should be of little effect owing to the reduced penetrability for angular momentum different from zero. Notice that the $^2P_{\frac{1}{2}} \rightarrow ^2P_{\frac{1}{2}}$ transition requires spin-orbit coupling; this is because the $J = \frac{3}{2}$ component of the total doublet p -wave must be excluded in order to preserve isotropy.

If now only the $^2S \rightarrow ^2S$ transition is taken into account, the statistical weight factor is $\omega = \frac{1}{3}$ and the maximum possible cross-section would be $\frac{1}{3}\pi\lambda^2$. At 200 keV this quantity is 2.3 barns, whereas the measured cross-section is 5.9 barns. One is therefore forced to the conclusion that strong spin-orbit coupling is present and that p -wave transitions may also be required to account for the magnitude of the cross-section. Admitting also the $^4S \rightarrow ^2D$ transition allows us to set $\omega = 1$ for the total s -wave reaction, thereby obtaining $\pi\lambda^2$, or 6.9 barns, for the total available cross-section when GQ is unity for both transitions. This is sufficient for our purpose. The symmetrical p -wave transitions may also give a sufficiently large cross-section, each having $(2l+1)\omega = \frac{1}{3}$; but the penetrability is here so much reduced that it is doubtful whether one could obtain values of GQ even approaching unity, particularly when we bear in mind that in addition spin-orbit coupling is required for both transitions.

Further difficulties appear when one tries to estimate the penetrabilities on the Gamow theory. In order to obtain sufficiently large penetrabilities one has to assume a reaction radius of the order of 15×10^{-13} cm. This point is discussed in more detail by Allan & Poole (1950).

The objection could be raised that s -waves should not contribute to the reaction at all since this would imply in both 2S and 4S states the close approach of identical

nucleons with the same spins, thereby violating the Pauli principle. The omission of the quintet s -state in the $D + D$ reaction by Konopinski & Teller (1948) is based on just this argument.* On the other hand, calculations by Tyrrell (1939) on the virtual states of ${}^5\text{He}$, the compound nucleus of the $T + D$ reaction, have shown that a 2S level should exist close to the well-known 2P doublet obtained by neutron scattering in helium. That is to say, a virtual 2S level of ${}^5\text{He}$ can exist at energies considerably lower than the ones we are dealing with (2P levels are at a neutron energy of 1 MeV, whereas the $T + D$ reaction produces neutrons of 14 MeV). Attempts to analyze the neutron scattering experiments in helium by Hall & Koontz (1947) in terms of the P -doublet have not been entirely satisfactory, and the presence of an S level is at least suggested by these experiments. If we are prepared to allow a 2S level at low excitation energies, there is no reason to exclude s -waves giving higher excitation energies. In general, it may be said that although the Pauli principle can preclude the formation of a bound s -state of low energy it does not rule out the possibility of s -wave transitions; at best it reduces their *a priori* probability. It is important in this connexion to bear in mind the difference between states of relative motion of systems of nucleons (i.e. nuclei) and states of motion of individual nucleons.

We conclude that the reaction can probably be described in terms of the transitions

$${}^2S \rightarrow {}^2S, \quad {}^4S \rightarrow {}^2D,$$

and it is attempted in this paper to explain the whole cross-section in terms of s -waves.

3. CROSS-SECTIONS

We may conveniently write the wave-function ψ describing the relative motion of triton and deuteron in the form adopted by Feshbach, Peaslee & Weisskopf (1947):

$$\phi(r) = r\psi = u + \eta v, \quad (4)$$

where the two functions u, v are those solutions of the Schrödinger equation which behave respectively as e^{-ikr} and e^{+ikr} at large distances. If $|\eta| = 1$ the wave-function describes simply a scattering process; if $|\eta| < 1$, it describes a process in which there is also absorption of the incoming particles. In no case may $|\eta|$ exceed unity.

The partial reaction cross-sections are then given by

$$\sigma_{\text{abs.}} = \omega\pi\lambda^2(1 - |\eta|^2), \quad (5)$$

and the scattering cross-sections by

$$\sigma_{\text{scatt.}} = \omega\pi\lambda^2 |1 + \eta|^2. \quad (5a)$$

Here we shall set $\omega = \frac{1}{3}$ for the 2S transition and $\frac{2}{3}$ for the 4S transition; in general, a different value of η will be required in each case. (Strictly speaking, $\sigma_{\text{abs.}}$ includes all inelastic processes. Since, however, the reaction (2) is the only one which occurs, $\sigma_{\text{abs.}}$ is the cross-section for this process.)

* This state can, however, be ruled out on quite different grounds. The only possible transition allowed by conservation of parity and total angular momentum is ${}^5S \rightarrow {}^1D_2$. This involves a change of spin of 2 units, thereby rendering it *a priori* very improbable.

We now express η in terms of the logarithmic derivative of the wave-function at the reaction radius c as defined in § 1:

$$[\phi(r)/\phi'(r)]_{r=c} = a + ib, \quad (6)$$

where a, b are of the dimension of length. This quantity we call the (complex) *reaction length*. Then we have, from (4),

$$\eta = \left[\frac{(a + ib) u'(r) - u(r)}{(a + ib) v'(r) - v(r)} \right]_{r=c}. \quad (7)$$

The functions u, v are obtained by selecting the appropriate solutions of the Schrödinger equation for the Coulomb field which, for zero angular momentum, takes the form

$$\left\{ \frac{d^2}{dr^2} + \frac{2}{r} \frac{d}{dr} + k^2 - \frac{1}{rd} \right\} \psi = 0. \quad (8)$$

Here d is the 'Bohr radius' and k is the reciprocal of λ . This is a confluent hypergeometric equation the solutions of which are of the type $W_{k,m}(z)$ of Whittaker & Watson (1920). It has been found more convenient in this paper to develop the solutions *ab initio* rather than to employ these functions explicitly.

If we knew sufficient about nuclear forces we should be able to write down and solve the Schrödinger equation of the system for values of r less than the reaction radius. We should then say that

$$a + ib = [\Phi/\Phi']_{r=c},$$

where Φ is the solution for this inner region. In the absence of such knowledge we have to make plausible assumptions about the expected dependence of $a + ib$ upon the bombarding energy. To a first approximation we may say that $a + ib$ is independent of energy. This is an assumption which can only be justified by application. A crude argument, however, would run as follows.

In a one-particle model the wave-function for the inner region is described by a periodic function

$$\Phi \sim \sin Kr,$$

where K is a wave-number. (We have here replaced the nucleus by a potential well.) Then since the logarithmic derivative must be continuous across any boundary, we have

$$[\phi/\phi']_{r=c} = [\Phi/\Phi']_{r=c} = \tan Kc/K.$$

K is the wave-number corresponding to the kinetic energy of the particle within the nucleus, which is large compared to the external kinetic energy and may therefore be regarded as practically independent of the latter. Provided, then, that Kc is not too near an odd multiple of $\frac{1}{2}\pi$, the reaction length will be independent of energy. If we add to the Hamiltonian for the compound system a term which allows the compound state to decay exponentially with time, representing absorption, we find that the reaction length should indeed be complex, its imaginary component representing the absorption width.

We shall find that the assumption of a constant reaction length is sufficient to explain qualitatively the energy dependence of the T + D cross-section: quantitative agreement over the whole measured energy range requires rather closer consideration.

4. SOLUTION OF THE WAVE-EQUATION

Since we are interested in $r\psi$ rather than ψ itself it is convenient to transform equation (8) by substituting $\xi = r\psi$. Then we obtain

$$\xi'' + \left(k^2 - \frac{1}{rd}\right)\xi(r) = 0. \quad (8a)$$

The general solution to this equation is proportional to

$$\xi(r) = f(r) + (\log r + \lambda)\xi_1(r), \quad (9)$$

where $\xi_1(r)$ is the regular solution given by

$$\left. \begin{aligned} \xi_1(r) &= r(1 + a_1 r + a_2 r^2 + \dots), \\ f(r) &= f_0 + f_1 r + f_2 r^2 + \dots, \end{aligned} \right\} \quad (10)$$

and

and the coefficients a_n, f_n satisfy the recurrence relations

$$\left. \begin{aligned} n(n+1)a_n &= a_{n-1}/d - k^2 a_{n-2}, \\ n(n+1)f_{n+1} &= f_n/d - k^2 f_{n-1} - (2n+1)a_n, \end{aligned} \right\} \quad (10a)$$

together with $a_0 = 1$, $a_1 = \frac{1}{2}d^{-1}$. f_1 is not determined by these relations; we choose $f_1 = 0$ for definiteness.

The general solution (9) contains in addition the parameter λ which has to be chosen in such a way as to make ξ behave in the required manner at infinity. The asymptotic behaviour of the wave-function can be studied by the method of Laplace transforms which yields the result that, if ξ is to behave at infinity as e^{+ikr} , it is necessary to set

$$\lambda = \log 2k + A(\alpha) - 1 + iP/2\alpha,$$

where

$$\alpha = (2kd)^{-1} = z_1 z_2 e^2/\hbar v,$$

v being the relative velocity;

$$A(\alpha) = \gamma + \sum_{n=1}^{\infty} \frac{\alpha^2}{n(n^2 + \alpha^2)},$$

where γ is Euler's constant ($= .5772 \dots$) and the series may be readily evaluated;

and

$$P(\alpha) = \frac{2\pi\alpha}{e^{2\pi\alpha} - 1}.$$

P represents an effective penetrability to zero distance, which may be seen by comparing solutions to the wave equation with and without the Coulomb term (Yost, Wheeler & Breit 1936). Notice that for low energies, i.e. large α , P reduces essentially to the Gamow factor

$$e^{-2\pi e^2/\hbar v}.$$

The corresponding wave-function is then given by

$$\xi(r) = f(r) + [\log 2kr - 1 + A(\alpha) + iP/2\alpha]\xi_1(r),$$

where $f(r)$ and $\xi_1(r)$ are defined by the equations (10) and (10a). This is the function behaving asymptotically as e^{ikr} ; the function which behaves as e^{-ikr} can be obtained by taking the complex conjugate. Hence, in the notation of equation (4), §3, we have

$$u = \xi^*(r) \quad \text{and} \quad v = \xi(r),$$

the asterisk denoting complex conjugate.

It is convenient at this point to separate ξ into its real and imaginary parts, p and q respectively:

$$\xi(r) = p(r) + iq(r).$$

According to equation (7) we then have

$$\eta = \frac{(ap' + bq' - p) + i(bp' - aq' + q)}{(ap' - bq' - p) + i(bp' + aq' - q)},$$

where p', q' are the derivatives of p, q with respect to r and the whole expression is to be evaluated at the point $r = c$, the reaction radius.

We require the following relation for the Wronskian of the two functions p, q which, according to equation (8a), must be independent of r :

$$pq' - p'q = Pd/2\alpha.$$

$$\text{Then} \quad 1 - |\eta|^2 = \frac{2bdP/\alpha}{(a^2 + b^2)(p'^2 + q'^2) + (p^2 + q^2) - 2a(pp' + qq') + bdP/\alpha} \quad (11)$$

$$\text{and} \quad |1 + \eta|^2 = \frac{4[(a^2 + b^2)p'^2 + p^2 + 2app']}{(a^2 + b^2)(p'^2 + q'^2) + (p^2 + q^2) - 2a(pp' + qq') + bdP/\alpha}. \quad (11a)$$

These expressions take the place of GQ of equation (1). They are expressed entirely in terms of calculated quantities and the two unknown parameters a and b .

5. DETERMINATION OF THE REACTION LENGTH

All the functions occurring in equations (11) can be computed in a straightforward manner. The series in r which occur in p and q need in general only the first three or four terms taking into account.

It remains to choose values of $a + ib$ to fit the observed cross-section as nearly as possible. As mentioned above, we should consider the 2S and 4S transitions separately, as these cannot interfere with each other. To each transition we therefore have to assign values of a and b , giving us in all four unknown parameters. In order to see what type of energy variation we may expect from equation (11) we shall for the time being assume that $a + ib$ is the same for both transitions. This assumption serves temporarily to dispose of two unknown parameters, thereby exhibiting more clearly the nature of the problem. Accordingly, we set the statistical weight factor ω equal to unity for the total s -wave transition. The values of a and b which we now set out to find will be some sort of average values for the two transitions.

If we divide the experimental cross-section at a single energy by $\pi\lambda^2$ we obtain the corresponding experimental value for $1 - |\eta|^2$. Equation (11) is then a quadratic equation for b , given a , or for a , given b . We impose the condition that a and b are both real; this condition determines both a and b within quite narrow limits. For example, from the 200 keV point we obtain from this procedure

$$-2.1 < a/d < -0.9, \quad 0.35 < b/d < 1.6.$$

Similarly, we obtain permissible domains of $a + ib$ for all energies; these are all found to overlap the 200 keV point completely. Thus the reaction length is found

to lie in the second quadrant of the complex plane, and its absolute magnitude is of the order of magnitude of the natural unit of length of the system. This seems physically reasonable.

In order to fit the observations as closely as possible it is convenient to consider the quantity

$$Z = \frac{\sigma}{\pi\lambda^2} \frac{2\alpha}{P} = \frac{1 - |\eta|^2}{P/2\alpha}, \quad (12)$$

which now contains only comparatively slowly varying functions of the energy and the lengths a and b . (Z is rather similar to an 'intrinsic reaction probability', the quantity assumed to be independent of energy and denoted by $|\alpha|^2$ in Konopinski & Teller's paper on the $D + D$ reaction (1948).) The experimental values of Z are shown in figure 1; they differ markedly from constancy. Also shown in this figure is a typical

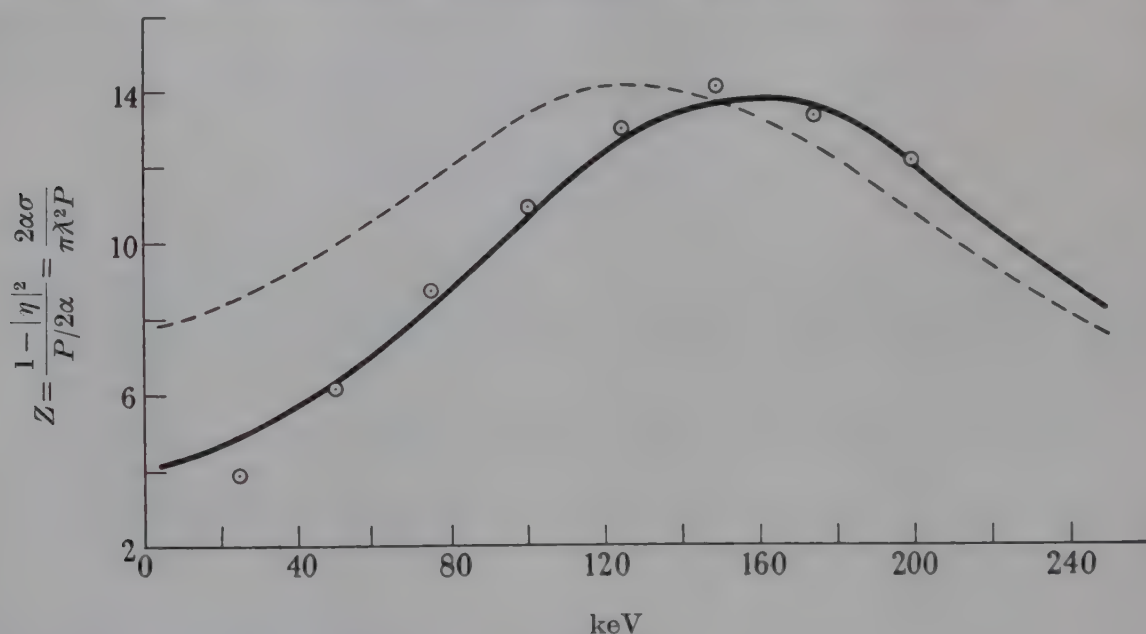


FIGURE 1. Intrinsic reaction probability: ○ experimental points; --- $a + ib$ constant, adjusted to fit at 150 keV; — $a + ib$ slowly varying according to the one-particle model.

curve obtained by taking a and b to be independent of energy. It is clear that the general features of the cross-section are reproduced in this approximation, but it is not possible to obtain quantitative agreement at all energies with any simple pair of values of a and b . Values of a and b are required at the low energy end which are numerically larger than those at higher energies. Moreover, this conclusion is not altered when we attempt to assign different values of a and b to the 2S and 4S transitions.

It is, of course, impossible to predict with any certainty how $a + ib$ should depend upon the energy. Nevertheless, it is tempting to see how far the one-particle model will take us towards getting quantitative agreement.

We proceed, as in §3, by writing

$$[\Phi/\Phi']_{r=c} = \tan Kc/K,$$

in the absence of absorption, where K is given by

$$K = \sqrt{\left(\frac{2M}{\hbar^2}\right)(E + W_0)^{\frac{1}{2}}}.$$

Here E is the bombarding energy, W_0 the kinetic energy inside the compound nucleus at zero bombarding energy, i.e. the depth of the well, and M the reduced mass. To take into account absorption we must include with W_0 the imaginary term $\frac{1}{2}i\Gamma$ where Γ is some absorption width, assumed to be small compared to W_0 . Then, to sufficient approximation,

$$a + ib = \frac{\tan Kc}{K} + \frac{i\Gamma c}{W_0} \sec^2 Kc. \quad (13)$$

The real and imaginary parts of the reaction length are now expressed in terms of E and the unknown, but assumed constant, quantities W_0 and Γ , in effect, we have replaced a and b by two other parameters, but we have not introduced a greater number of parameters.

In order that a and b shall show appreciable variation it is necessary that Kc should lie near an odd multiple of $\frac{1}{2}\pi$. Furthermore, since we know that $a + ib$ lies in the second quadrant, we can set

$$Kc = (2n + 1)\frac{1}{2}\pi + \theta,$$

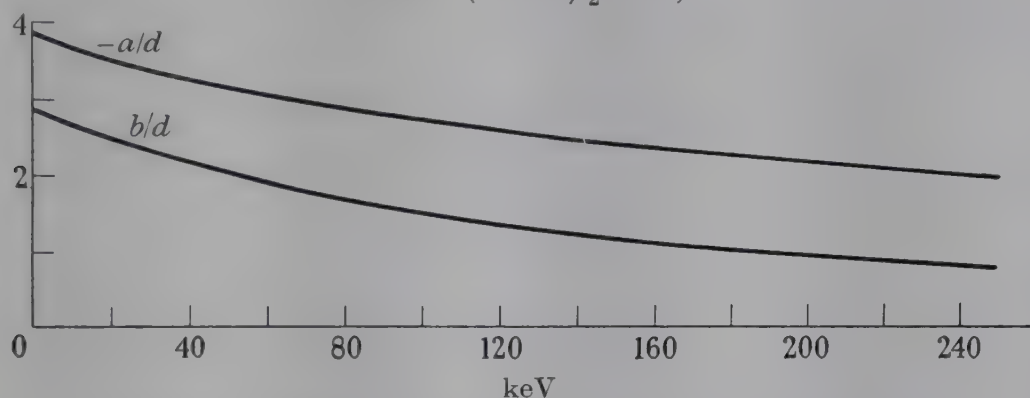


FIGURE 2. Variation in $a + ib$ given by the one-particle model.

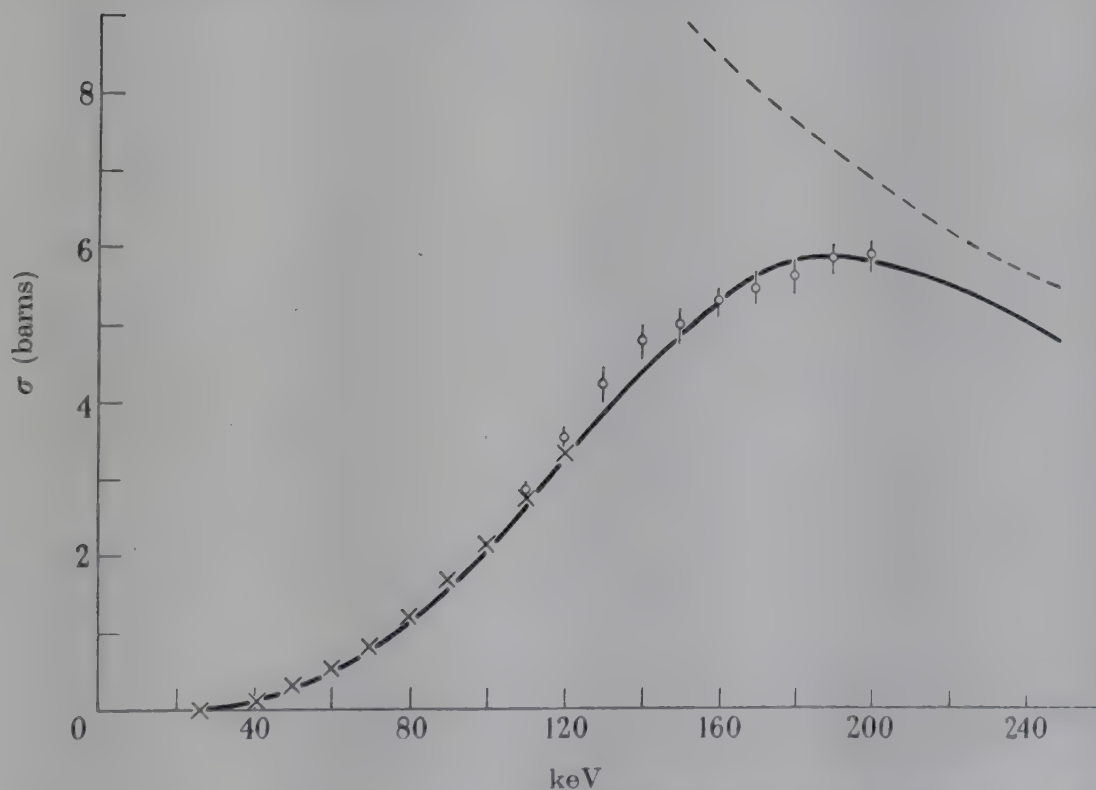


FIGURE 3. Reaction cross-section: --- calculated; $\times$ Bretscher & French; ϕ , Allan & Poole; --- $\pi\lambda^2$.

where θ is small and positive. This enables us to choose n so that W_0 is of the order of magnitude of nuclear kinetic energies. Both a and b then decrease numerically with increasing energy which is just the behaviour we require.

The best fit to the cross-section obtained by the one-particle model is shown in figures 1 and 3. The agreement is surprisingly good, and perhaps fortuitously so, particularly in that we have assumed $a + ib$ to be the same for both transitions. At energies below 25 keV this approximation predicts somewhat larger cross-sections than are observed. In general it may therefore be said that the variation in a and b predicted by equation (13) seems to hold good provided that it is a slow variation with energy. At the lowest energies (13) varies more rapidly and the approximation breaks down.

6. DISCUSSION

It has been shown that the whole of the T + D cross-section may be explained on the assumptions that it is due to s -waves only, and that it can be described by a complex reaction length which varies slowly with energy. Expressed in terms of the Bohr radius $d = 12.2 \times 10^{-13}$ cm. we have taken the reaction radius to be $c = 0.57d$, and we have found a reaction length of the same order of magnitude.

We have assumed that the 2S and 4S transitions behave in the same manner, and have found that the process can be explained adequately in terms of the remaining two parameters. Had we considered the two transitions to behave differently we should have had two further parameters to adjust, or four in all. Information regarding the 4S transition itself (as also the P transitions) cannot be extracted from experiments hitherto performed; experiments performed with polarized nuclei are required for this purpose. Until such experiments have been done, there seems little likelihood of obtaining independent reaction lengths for the different transitions.

The radius c of the reaction surface has not been regarded as a variable parameter because the ratio ξ/ξ' , and therefore η , depends rather insensitively upon it. We have taken c as fixed and have found a and b to suit observation. Had we taken a slightly different value of c , values of a and b would have been found which suited equally well. We are therefore not acutely concerned with the precise numerical definition of the quantity c . This is not the case with the Gamow theory where the energy dependence is a much stronger function of c . Nor is it the case in the present theory for angular momentum different from zero, η then depending upon c^{2l+1} , although it is still better than the corresponding Gamow theory. For this reason we have made no attempt to estimate at this stage the small amount of p -wave interaction which may be present. Introduction of p -waves would also mean extra undetermined parameters which we have found unnecessary for an adequate description of the process. We cannot, however, exclude the possibility of symmetrical p -wave transitions.

It is submitted that the function Z defined by equation (12) is the proper one to investigate in considering light nuclear reactions with charged particles. It contains in intimate union the effects of the specifically nuclear forces through the quantities a and b , and the effects of the Coulomb field through the wave-functions p and q . By consideration of this quantity we conclude that the energy dependence of the T + D

cross-section is primarily due to the behaviour of the exact wave-functions of the Coulomb field, and that previous attempts to explain it have been unsuccessful owing to the failure of the WKB approximation. Irrespective of whether we introduce p -waves or not, we have to assume the presence of strong spin-orbit coupling to explain that part of the cross-section which exceeds that arising from the pure ${}^2S \rightarrow {}^2S$ transition.

I wish to thank Professor Pryce for suggesting the problem and for his many helpful comments. Thanks are also due to Dr E. Bretscher and Mr M. J. Poole for discussions of their experiments. I would like to take this opportunity of expressing gratitude to Dr O. Buneman and Dr B. Davison for many helpful discussions on this and other topics during the past year. This work is published by kind permission of the Director, Atomic Energy Research Establishment.

REFERENCES

- Allan, D. L. & Poole, M. J. 1949 *Nature*, **164**, 102.
Allan, D. L. & Poole, M. J. 1951 *Proc. Roy. Soc. A*, **204**, 488, 500.
Bethe, H. 1937 *Rev. Mod. Phys.* **9**, 2.
Bretscher, E. & French, A. P. 1949 *Phys. Rev.* **75**, 8.
Fermi, E. & Marshall, L. 1947 *Phys. Rev.* **71**, 66.
Feshbach, H., Peaslee, D. C. & Weisskopf, V. F. 1947 *Phys. Rev.* **71**, 145.
Hall, T. A. & Koontz, P. G. 1948 *Phys. Rev.* **72**, 196.
Konopinski, E. J. & Teller, E. 1948 *Phys. Rev.* **73**, 822.
Tyrrell, W. A. 1939 *Phys. Rev.* **56**, 250.
Whittaker, E. T. & Watson, G. N. 1920 *Modern analysis*, 3rd ed. p. 339. Cambridge University Press.
Yost, F. L., Wheeler, J. A. & Breit, G. 1936 *Phys. Rev.* **49**, 174.

Orientation of fatty acid and soap films on metal surfaces

BY J. W. MENTER AND D. TABOR

*Research Laboratory on the Physics and Chemistry of Rubbing Solids,
Department of Physical Chemistry, University of Cambridge*

(Communicated by F. P. Bowden, F.R.S.—Received 9 June 1950)

[Plates 12 and 13]

An electron-diffraction study has been made of the structure and orientation of fatty acid films and soap films adsorbed on metal surfaces. In particular, an examination has been made of the effect of temperature on the orientation, using the reflexion method and an improved experimental technique. The results show, as previous workers have found, that at room temperature the first layer is generally oriented with the hydrocarbon chains normal to the surface, though other orientations may be obtained by 'rubbing' the film. As the temperature is raised the 'layer lines' become increasingly less distinct, and at a characteristic temperature only the pattern of the substrate is observed. This transition is a true disorientation effect, since on cooling orientation is again observed. At higher temperatures the film may evaporate completely from the surface.

With fatty acids on non-reactive metals such as platinum the disorientation temperature is close to the bulk melting-point of the fatty acid. On reactive metals such as zinc and cadmium the disorientation occurs at higher temperature which is close to the bulk melting-point of the appropriate metallic soap suggesting the formation of the soap by chemical reaction. This is supported by the similar behaviour of thin films of metallic soap deposited directly on the metal substrate. These observations are compared with the lubricating properties of fatty acids and soaps on metal surfaces.

INTRODUCTION

Several workers (for example, Andrew 1936; Finch & Zahoorbux 1937; Germer & Storks 1938; Havinga & de Wael 1937, 1940; Trillat & Motz 1935) have studied by electron diffraction the orientation at room temperature of thin films of long-chain carbon compounds on metal surfaces. More recently Tanaka (1938, 1939, 1941), Brummage (1947 *a, b*) and Cowley (1948) have extended the investigation to the variation with temperature of the structure of these films. In particular, Brummage has established that there is a transition temperature characteristic of the chemical nature of the surface and of the deposited film, at which such films lose their characteristic orientation. Most of his results, however, refer to relatively thick films deposited by evaporation from solution. The present investigation is concerned largely with the behaviour of films of the order of one molecule in thickness adsorbed from the melt or formed by a standardized rubbing method. The structure of these films and the variation of structure with temperature has been investigated using an improved experimental technique, and the results provide direct information about the properties of thin films in immediate contact with the metal substrate.

Boundary lubrication is effected by lubricant layers of comparable thickness, and a comparison is made of the electron-diffraction results and the effect of temperature on the lubricating properties of the same compounds.

*a**b**c**d**e**f*

FIGURE 1

a, myristic acid*b*, palmitic acid*c*, stearic acid*d*, octacosanoic acid

} effect of chain length on orientation (retracted layer on zinc)

e, electron beam perpendicular to rubbing direction*f*, electron beam parallel to rubbing direction} rubbed layer of zinc octacosanoate
on zinc.

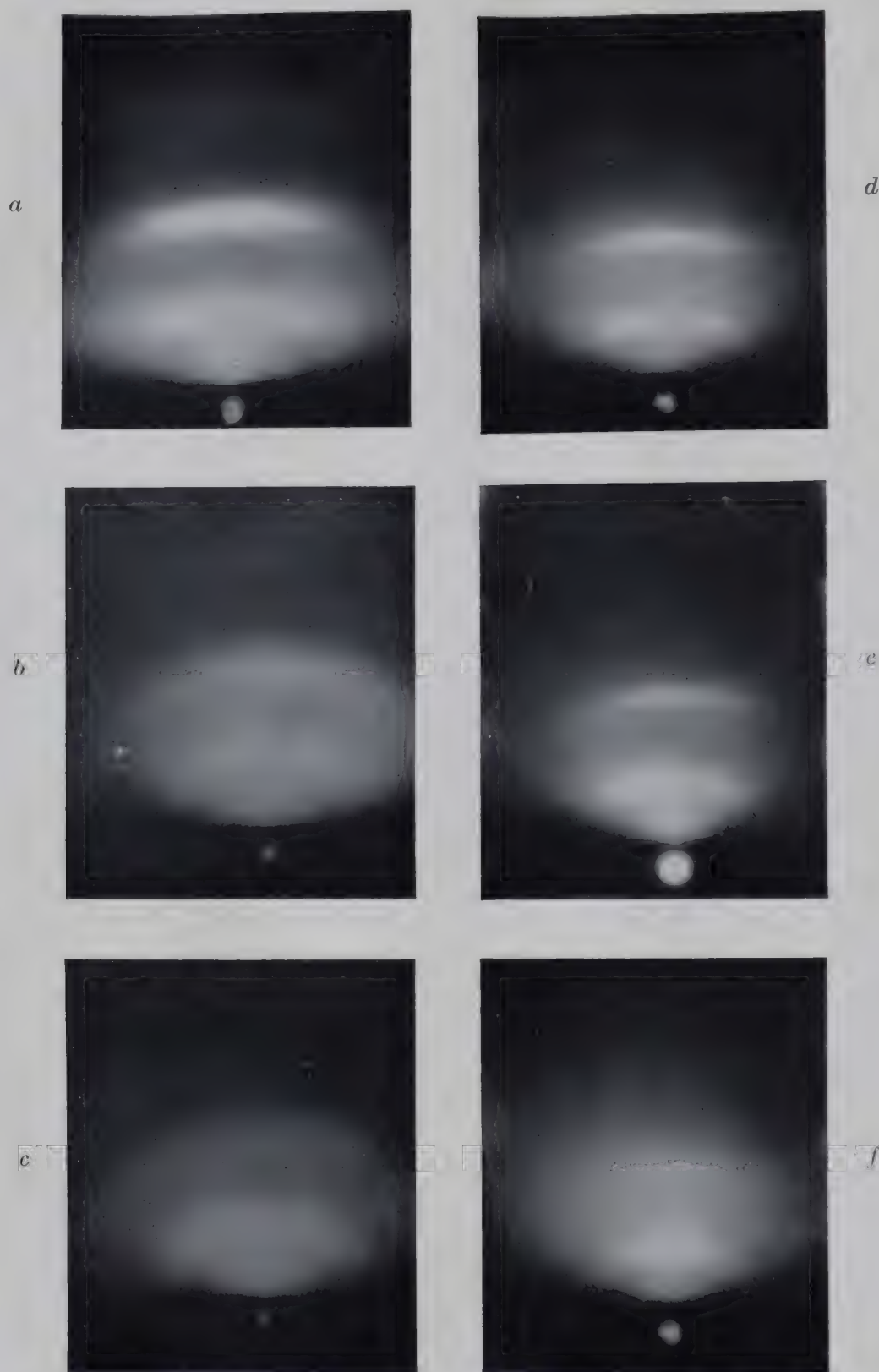


FIGURE 2. *a* (22°C), *b* (47°C), *c* (66°C), retracted layer of stearic acid on platinum; *d* (23°C), *e* (68°C), *f* (88°C), retracted layer of stearic acid on zinc.

Apparatus

A cold-cathode camera of the Finch type, working in the range 35 to 40 kV, was used.

The specimen holder specially designed for heating specimens is described elsewhere (Menter & Sanders 1951). The metal specimen is in the form of a flat disk $\frac{3}{4}$ in. in diameter, and may be heated in the vacuum to temperatures of the order of 200° C. The surface temperature is measured by a thermocouple and may be determined to within 2° C.

PREPARATION OF REFLEXION SPECIMENS AND EXPERIMENTAL PROCEDURE

The platinum, silver, copper and nickel specimens were spectroscopically pure, the mild steel was taken from a bar of undetermined composition. The zinc and cadmium were prepared by casting from chemically pure samples.

A standard procedure was adopted for the preparation of the metal surface. After preliminary grinding under water with 600 emery, it was polished under alcohol on a polishing cloth impregnated with alumina and degreased electrolytically according to a procedure recommended by Blodgett (1935). The specimen was then washed in cold tap water and boiling distilled water. Finally, it was cleaned by blowing off the hot water and projecting a stream of hot air on to the surface for about 30 seconds.

With all metals except platinum and silver the pattern obtained from the clean surface by electron diffraction was that of the oxide.

Among the methods available for depositing the acid or soap layers, those investigated were (1) the retraction method, (2) the rubbing method and (3) the Langmuir-Blodgett method. In method (1) a few crystals of the acid are placed on the cleaned surface and melted on a hot plate. By spreading the molten acid with a glass rod the surface may be covered by an adsorbed layer of molecules. This is not wetted by the bulk liquid which therefore retracts into a number of droplets. These are removed by soaking off with a filter-paper. In method (2) the crystals are melted as before and the liquid is spread over the surface. The bulk is soaked off with filter-paper, no great care being taken to remove it all. The surface is allowed to cool and then rubbed vigorously, generally in one direction only, on filter-paper. With the metal soaps this method is modified slightly. The soap is not melted, but rubbed on to the metal surface cold. A few experiments with the Langmuir-Blodgett dipping method showed that the layers recrystallized very rapidly on heating, giving variable results. This method was not pursued further.

The patterns obtained from very thin oriented layers of long-chain carbon compounds are rather diffuse and become progressively less distinct as the temperature is raised. It was therefore considered undesirable to rely on visual observation of the fluorescent screen as the temperature was raised. A film-holder was therefore designed with which it was possible to make up to fifteen or twenty exposures without breaking the vacuum (see Menter & Sanders 1951).

A photograph was taken at room temperature, care being taken that the beam did not strike any one part of the specimen for an unduly long time, in order to

reduce local heating to a minimum. The specimen was then heated, the rate of heating being of the order of 4°C per minute. Photographs were taken when major changes in the nature of the pattern occurred and also at intervals of about 5°C in the region of the transition temperature. The transition temperatures quoted in the text are accurate to within 5°C .

EXPERIMENTAL RESULTS

Retracted layers of acids

Layers of the following acids were examined on platinum, silver, zinc, cadmium, copper, nickel and mild steel surfaces: caprylic acid (m.p. 16.35°C), capric acid (m.p. 30.5°C), lauric acid (m.p. 44.0°C), myristic acid (m.p. 53.6°C), palmitic acid (m.p. 62.7°C), stearic acid (m.p. 69.2°C), octacosanoic acid (m.p. 90.4°C).

Layers of straight-chain fatty acids deposited by retraction give a pattern which consists of a set of diffuse layer lines parallel to the shadow edge, the spacings of which correspond to successive orders of diffraction from the alternate C-C spacing in the hydrocarbon chain. The pattern is unchanged when the specimen is rotated about an axis perpendicular to its surface, showing that the molecules are oriented with their hydrocarbon chains approximately normal to the surface. Well-oriented patterns show an intensity minimum in the centre of the odd orders of diffraction, of which that in the first order is most readily visible. If the hydrocarbon chains are disposed at a small angle to the general surface normal, the azimuthal angle of the chains being random, then the layer lines are slightly curved. The greater the departure from the surface normal, the greater the curvature of the layer lines. For a fuller discussion see Murison (1934), Karle (1946), Karle & Brockway (1947). Very well-oriented layers, in which the departure from the surface normal is no greater than a few degrees, are formed only by acids above palmitic (C_{16}) in the series. Below this chain length, the orientation becomes increasingly poorer, and is not observed below lauric acid (C_{12}). The change in orientation with chain length is shown for a series of acids on zinc in figure 1*a*, *b*, *c* and *d*, plate 12.

If the retraction of the molten liquid is not perfect there may be a few crystallites of the acid lying in random azimuth with their hydrocarbon chains parallel to the surface. The pattern from these crystallites consists of sharp rings or arcs of small radius, corresponding to the so-called 'short spacings' of the acids.

On heating the specimen the pattern maintains the same appearance as at room temperature up to a temperature near the transition. The layer lines then begin to curve, showing that the molecules are starting to undergo relatively large thermal oscillations; the intensity minimum in the centre of the first-order diffraction band and the higher orders of the pattern disappear; the remaining two orders rapidly take on the appearance of two haloes which gradually disappear and reveal the pattern of the substrate oxide or metal. The transition temperature is taken to be that at which the layer lines disappear, leaving only a diffuse background of incoherently scattered electrons.

The transition temperature is markedly affected by the chemical nature of the surface (see table 1 and figure 3). Thus on platinum the temperature is close to the

bulk melting-point of the acid, whereas on zinc, cadmium, nickel or mild steel it is higher, for instance 20 to 30° C for stearic acid and 40° C for octacosanoic acid. This behaviour is illustrated in figure 2, plate 13, by a series of photographs of a stearic acid layer on platinum and zinc respectively. Copper and silver show intermediate behaviour, the shorter chain-length transition temperature being high compared with that on platinum, whereas the temperature for longer chains becomes almost equal to that on platinum. The metals thus fall into three groups as shown in figure 3.

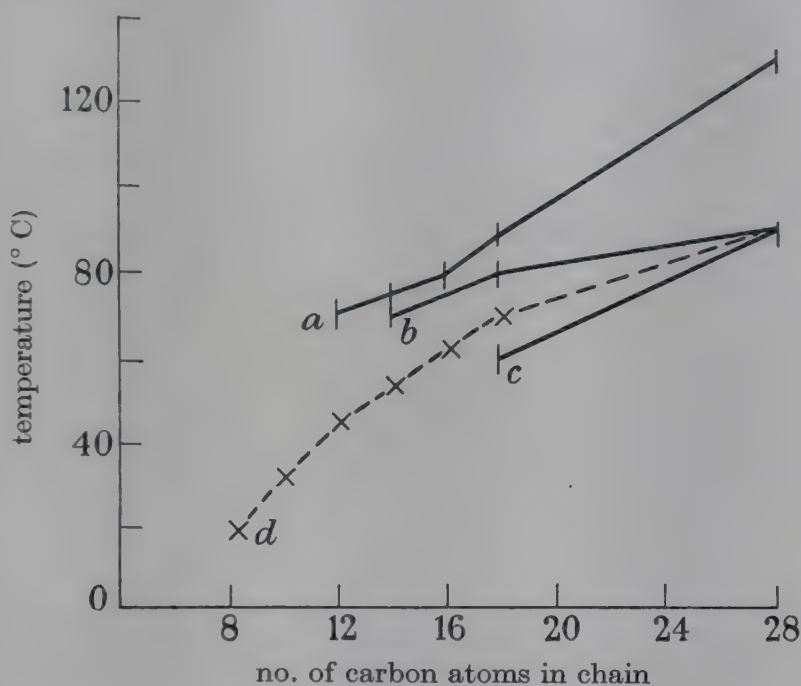


FIGURE 3. Transition temperatures for retracted layers of fatty acids: *a*, zinc, cadmium; *b*, copper; *c*, platinum; *d*, bulk melting-point of fatty acid.

Rubbed layers of fatty acids

The general trend of the results is similar to that described above except that the shorter chain molecules are better oriented when rubbed than when deposited by retraction. There may also be small crystallites lying on top of the normally oriented layer, showing a preferred orientation.

The changes observed in the pattern on heating are similar to those reported for the retracted layers (see table 1 and figure 4), except that on reactive metals, cadmium, zinc, mild steel and nickel, the temperature is somewhat higher and is close to the softening point of the appropriate soap formed by reaction between the acid and the material of the surface. In addition, the transition appears to be less abrupt than with the retracted layers.

With octacosanoic acid the behaviour is different. Rubbing produces a marked tilting of the hydrocarbon chains along the direction of rubbing, the angle of inclination from the surface normal being of the order of 16°. On heating the specimen, the inclined orientation is maintained up to about 60° C; above this temperature the molecules gradually revert to the normal orientation, reaching it at a temperature close to the bulk melting-point of the acid. The final disorientation then occurs in the usual way.

With copper it is found that although a soap is formed it does not remain oriented normal to the surface up to the softening point of the soap. Near the melting-point of the acid it forms very small crystals oriented with the hydrocarbon

TABLE 1. TRANSITION TEMPERATURES OF LAYERS OF FATTY ACIDS AND SOAPS

metal	deposited layer acids	transition temperature (° C)		m.p. of acid (° C)	softening point of soap (° C)
		retracted layer	rubbed layer		
platinum	lauric (C ₁₂)	—	35	44·0	—
	palmitic (C ₁₆)	—	55	62·7	—
	stearic (C ₁₈)	60	60	69·2	—
	octacosanoic (C ₂₈)	90	80	90·4	—
silver	myristic (C ₁₄)	—	80	44·0	—
	stearic (C ₁₈)	—	85	69·2	180
copper	lauric (C ₁₂)	—	70	44·0	100
	stearic (C ₁₈)	80	80	69·2	111
	octacosanoic (C ₂₈)	90	90	90·4	125
mild steel	lauric (C ₁₂)	60	105	44·0	—
	stearic (C ₁₈)	95	125	69·2	115
	octacosanoic (C ₂₈)	125	140	90·4	—
zinc	lauric (C ₁₂)	70	110	44·0	120
	myristic (C ₁₄)	75	—	53·6	—
	palmitic (C ₁₆)	80	—	62·7	—
	stearic (C ₁₈)	90	130	69·2	—
nickel	octacosanoic (C ₂₈)	130	145	90·4	130
	lauric (C ₁₂)	—	85	44·0	—
	stearic (C ₁₈)	100	120	69·2	—
cadmium	octacosanoic (C ₂₈)	140	145	90·4	—
	lauric (C ₁₂)	—	95	44·0	—
	stearic (C ₁₈)	90	130	69·2	110
	octacosanoic (C ₂₈)	140	140	90·4	—
soaps					
platinum	cadmium stearate	—	95	—	110
cadmium			100	—	110
platinum	copper stearate	—	90	—	111
copper			90	—	111
platinum	copper octacosanoate	—	105	—	125
copper			105	—	125
platinum	zinc octacosanoate	—	135	—	130
zinc			150	—	130

chains parallel to the surface. The temperature at which this change occurs is given as the transition temperature in table 1. These small crystals persist up to the softening point of the soap.

Structure and behaviour of crystallites

The pattern from the crystallites lying above the normally oriented layer consists usually of a single sharp ring corresponding to a spacing of about 4 Å. This is the (110) ring of the fatty acids, which remains approximately constant in spacing

throughout the series. It arises from crystallites lying above the normally oriented layer with their hydrocarbon chains approximately parallel to the surface. Sometimes the ring shows axial arcing indicating that there is a preferred orientation with the (110) planes parallel to the surface. The ring is often diffuse, showing that the crystallites are very small. On heating, this pattern, with stearic acid, for instance, on all metals disappears quite suddenly about 45 to 50° C, well below the bulk melting-point of the acid, 69.2° C, giving place to a diffuse halo of about the same spacing. The pattern does not reappear on cooling after heating to the transition temperature of the normally oriented layer.

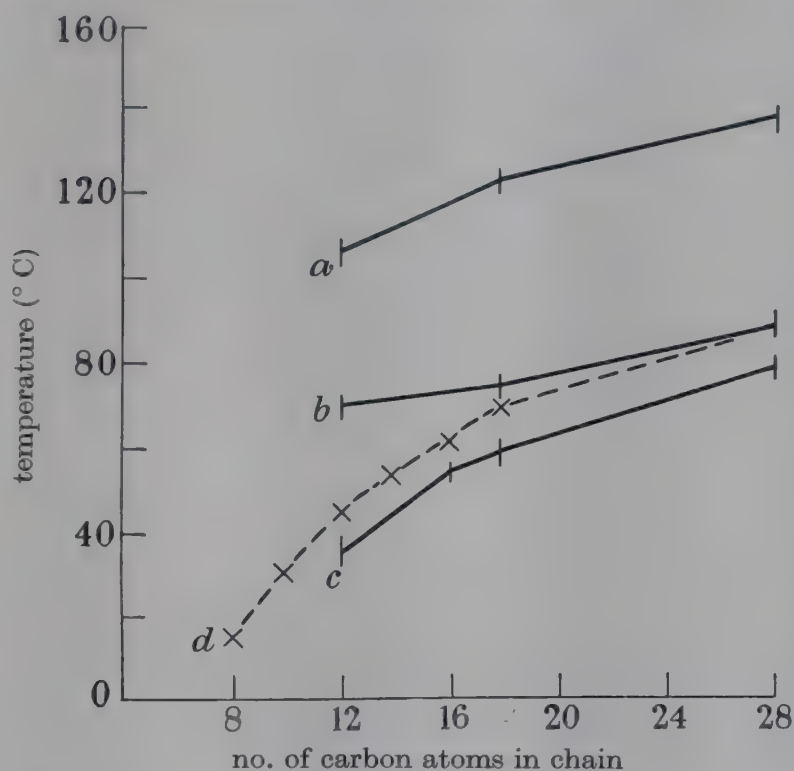


FIGURE 4. Transition temperatures of rubbed layers of fatty acids: *a*, zinc, cadmium, mild steel; *b*, copper; *c*, platinum; *d*, melting-point curve.

Metal soap layers

It is not possible to deposit the soaps by the retraction method used for the acids. The only satisfactory way of depositing thin layers was found to be that of rubbing the solid soap on the cold surface. In general, two experiments were performed with each soap. The first was on a surface of the parent metal from which the soap was derived and the second with the soap on a platinum surface.

At room temperature, fairly well-oriented layers of cadmium stearate and copper stearate may be formed with the chains normal to the surface. In cases where the parent metal and platinum were both used as the substrate metal, there appeared to be no difference between the orientation of the layers on the different metals. In general, the layers are not so well oriented as are the corresponding acid layers on the reactive metals. On heating, the final disorientation occurs at 10 to 20° C below the bulk melting-point of the soap, the general behaviour being similar to that described for the acids (see table 1).

Soaps derived from octacosanoic acid show a behaviour similar to that observed with the acid on the parent metal. The hydrocarbon chains are inclined to the surface (see figure 1*e, f*), and on heating the inclination diminishes, becoming zero at about 85° C.

DISCUSSION

Structure at room temperature

It is supposed that when the molten acid is placed in contact with the surface, a closely packed layer of molecules is adsorbed, in the first place physically, as a result of the attraction between the dipole of the —COOH group and its image in the surface. This adsorbed layer presents a surface of CH₃ groups to the molten acid by which it is not wetted. It is uncertain whether the oriented layer follows the surface contours or consists of small aggregates at the surface asperities.

The layers formed by the retraction method are almost certainly monomolecular (Courtney-Pratt 1950). It is difficult, however, to estimate unambiguously the thickness of the layers formed by the rubbing method. The pattern arises from electrons scattered by hydrocarbon chains oriented on the surface asperities. At room temperature, particularly with the shorter chain acids, the rings from the metal substrate are sometimes visible simultaneously with the pattern from the hydrocarbon chains. Schoon (1938) has shown that for 25 kV electrons the depth of penetration in a direction normal to the general direction of the surface in the reflexion method is less than 20 Å. With 40 kV electrons it is somewhat greater than this. The length of the lauric acid chain is about 15 Å, so that a weak pattern from the substrate would be expected in this case; if the surface film is monomolecular on the other hand, it must be admitted that the surface film may be incomplete, and that the rings may arise from asperities not covered by hydrocarbon chains. Previous workers have assumed that a band pattern in which the bands show no structure in the form of spots along their length arises from a monolayer. This criterion is, however, not unambiguous. Such a pattern could equally well arise from a film consisting of a large number of small crystallites arranged randomly in azimuth but with their hydrocarbon chains aligned parallel to the surface normal.

Effect of heating

The changes in the diffraction pattern on heating show that the molecules undergo thermal oscillations of increasing amplitude. At the transition temperature one end of the molecule is still anchored to the surface, but the alignment of the hydrocarbon chains has disappeared completely. The dependence of the transition temperature on the chemical nature of the substrate distinguishes the acids from paraffins, alcohols and esters. Sanders has shown that for these compounds the transition temperature is independent of the nature of the substrate. This at once suggests that the high transition temperatures of the acids on the reactive metals is associated with the formation *in situ* of a metallic soap layer which is more firmly attached to the surface than the physically adsorbed acid. Direct evidence for the occurrence of chemical reaction is provided by the radioactive tracer experiments of Bowden & Moore (1949).

The experiments show that the layer is more firmly attached to reactive surfaces. For example, the platinum surface was completely hydrophilic after heating to 40 to 50° C above the transition temperature, whereas with zinc, cadmium, nickel or mild steel, the surface remained hydrophobic after the same treatment. This view is supported by the fact that the pattern from the platinum surface becomes quite clear after the disorientation of the film, whereas that of the reactive metals remains very diffuse and obscure. This is presumably due to the fact that in the latter case molecules of the surface film, although disoriented, remain attached to the surface by their dipolar heads and prevent penetration of electrons to the metal substrate.

That the disorientation is truly reversible has been checked by experiments on platinum and zinc. With platinum a well-oriented pattern may be obtained on cooling after heating to the transition temperature, but after heating to 40 to 50° C above the transition temperature no pattern reappears. This suggests that the layer has evaporated from the surface. With zinc, on the other hand, even after heating to 50° C above the transition temperature, a well-oriented pattern reappears on cooling which improves with time. However, the orientation of the layer at a given temperature during cooling is in general not so good as that at the same temperature during heating.

With acids on reactive metals or with the soaps themselves the diffraction bands begin to curve at a temperature close to the bulk melting-point of the parent acid. This suggests that the hydrocarbon chains have more freedom at a temperature well below the bulk melting-point of the soap. This behaviour is apparently connected with the 'genotypical' changes described by Thiessen *et al.* (1931, 1933, 1935, 1936).

The transition temperatures of the rubbed layers are slightly greater than those of the corresponding retracted layers. The reason for this is not clear, but it may be because of the slightly better alinement of the chains in the rubbed layer. The only other significant way in which layers formed by the two methods differ is in the behaviour of octacosanoic acid films. It seems possible that the sloping alined orientation of the rubbed layers is a direct mechanical effect produced by rubbing. The sloping orientation is unstable, and on heating the chains are able to regain the normal orientation on account of their increased thermal energy. It is significant, in connexion with the remarks above concerning a 'genotypical' change, that the normal orientation is assumed near the bulk melting-point of the acid, both for layers deposited originally as acid and for the soap directly applied to the surface.

The transition temperatures of soap films are of the same order as those observed with the corresponding fatty acid films deposited by retraction on the appropriate metals, but they are somewhat lower. This is not altogether unexpected, since the soap molecule formed *in situ* almost certainly has the metal atom in the plane of the surface with the hydrocarbon chains at a steep angle, whereas the soap film formed by applying the bulk soap may have a number of molecules with the hydrocarbon end next to the surface. Such molecules would not be so strongly adsorbed, and, moreover, might reduce the lateral cohesion of the film.

Correlation with friction measurements

It is interesting to compare these results with experiments on the boundary lubricating properties of fatty acids and soaps. With most lubricated metals it is found that, at room temperature, the friction is low, the surface damage small and the sliding smooth. As the temperature is raised little change occurs until a characteristic transition temperature is reached at which the friction and surface damage increase and the sliding becomes irregular (Bowden & Leben 1940; Bowden, Gregory & Tabor 1945). The effect is reversible and has been associated with a disorientation or desorption of the lubricant molecules (Tabor 1940, 1941; Hughes & Whittingham 1942; Frewing 1943).

In attempting a correlation between the friction and electron-diffraction transition temperatures the physical differences between the two experimental conditions must be considered. In the diffraction camera we have, in general, a system: metal—metal oxide—hydrocarbon film—vacuum; whereas in the friction measuring

TABLE 2. ELECTRON DIFFRACTION AND FRICTION TRANSITION TEMPERATURES OF FATTY ACIDS AND SOAPS

deposited layer	surface	transition temperature (° C)		softening point of soap (° C)
		electron diffraction	friction*	
lauric acid	platinum	35	40	—
	silver	50	45	—
	copper	70	{ 50 80	100
	mild steel	105	120	—
	cadmium	95	70	—
	zinc	110	120	120
stearic acid	platinum	60	65	—
	silver	85	77	180
	copper	80	{ 108 105	111
	mild steel	125	140	115
	cadmium	130	100	110
	zinc	130	—	—
octacosanoic acid	platinum	80	85	—
	copper	90	{ 110 130	125
	mild steel	140	—	—
	cadmium	140	{ 144 120	—
	zinc	145	130	130
	cadmium	100	130	110
cadmium stearate	platinum	95	—	—
	copper	90	110	111
copper stearate	platinum	90	—	—
	copper	105	120	125
copper octacosanoate	platinum	100	—	—
	zinc	150	130	130
zinc octacosanoate	platinum	135	120	—

* Friction results taken from Ballantyne (1945); Bowden, Gregory & Tabor (1945); Tabor (1941) and Shooter (1949).

apparatus we have: metal—metal oxide—hydrocarbon film—possibly bulk of lubricant—hydrocarbon film—metal oxide—metal. In addition, in the friction experiments there are high local pressures and shear stresses tending to disrupt the lubricant film; on the other hand, there is often an excess of lubricant ready to repair possible rupture of the film. Thus we cannot expect an exact identity between the transition temperatures found in the two types of experiment.

Collected data on friction transition temperatures are presented in table 2 together with transition temperatures of rubbed layers observed by electron diffraction. All friction measurements were carried out with an excess of lubricant

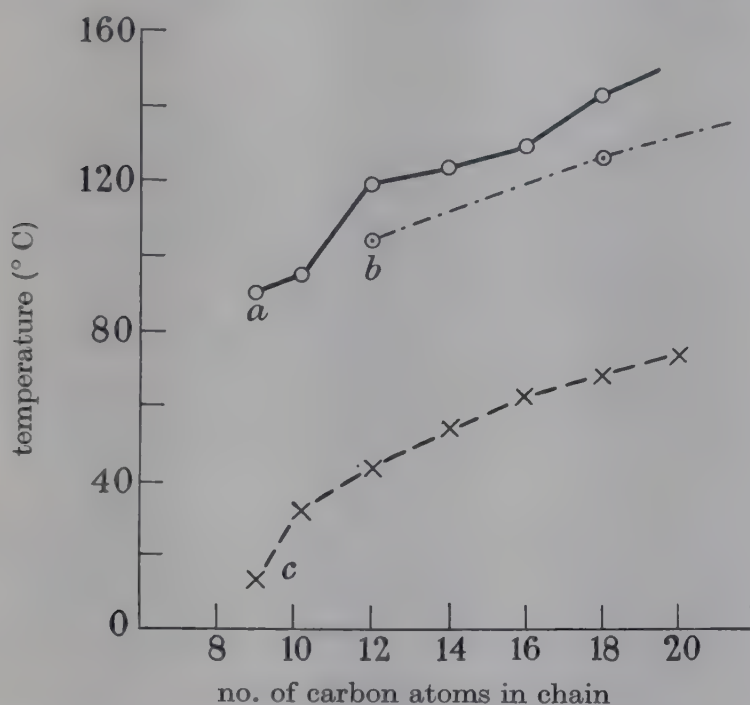


FIGURE 5. Transition temperatures for fatty acids on mild steel: *a*, friction; *b*, electron diffraction; *c*, bulk melting-point of fatty acid.

round the slider. In general, the agreement between the results is good. With the unreactive metal, platinum, the friction coefficient rises near the bulk melting-point. The electron-diffraction transition temperature for the acid has approximately the same value. On cadmium, zinc and mild steel which form soaps very readily, both temperatures are higher and near the softening point of the soap. With copper the friction transition temperature appears to be much higher. Thus the recrystallized layer, which is formed above the transition temperature (see p. 518), still provides some lubrication up to its melting-point.

It is interesting to compare the variation of the transition temperature with chain length in the two series of experiments. This is shown in figure 5 for mild steel. The friction results are those of Tabor (1941). It is clear that there is a very good correlation in this case, although it must be mentioned that other workers have found anomalously high friction transition temperatures with mild steel. Nevertheless, the results support the view of Bowden, Tabor and others that effective boundary lubrication is obtained only when there is strong *lateral* adhesion between the lubricant molecules. In this way metallic interaction between the

sliding surfaces is reduced to a very small value. On heating, the lateral adhesion is increasingly overcome by the thermal motion of the molecules and at a temperature which is close to the disorientation temperature of the lubricant film there is a marked increase in metallic interaction with a corresponding rise in the friction and the amount of surface damage.

We wish to thank Dr F. P. Bowden for advice and encouragement, Dr W. Cochrane and Dr J. V. Sanders for helpful discussions, the Ministry of Supply (Air) for a Research Grant and the D.S.I.R. for a maintenance grant to one of us (J.W.M.).

REFERENCES

- Andrew, L. T. 1936 *Trans. Faraday Soc.* **32**, 607.
Ballantyne, J. 1945 *Lubrication and friction report*, no. 51. C.S.I.R. (Aust.).
Blodgett, K. 1935 *J. Amer. Chem. Soc.* **57**, 1007.
Bowden, F. P., Gregory, J. N. & Tabor, D. 1945 *Nature*, **156**, 97.
Bowden, F. P. & Leben, L. 1940 *Phil. Trans. A*, **239**, 1.
Bowden, F. P. & Moore, A. C. 1949 *Research*, **2**, 585.
Brummage, K. G. 1947*a* *Proc. Roy. Soc. A*, **188**, 414.
Brummage, K. G. 1947*b* *Proc. Roy. Soc. A*, **191**, 243.
Courtney-Pratt, J. S. 1950 *Nature*, **165**, 346.
Cowley, J. M. 1948 *Trans. Faraday Soc.* **44**, 60.
Finch, G. I. & Zahoorbux, F. D. 1937 *Inst. Mech. Engrs, Discussion on Lubrication*, **2**, 295.
Frewing, J. 1943 *Proc. Roy. Soc. A*, **181**, 23.
Germer, L. H. & Storks, K. H. 1938 *J. Chem. Phys.* **6**, 280.
Havinga, E. & de Wael, J. 1937 *Rec. Trav. chim. Pays-Bas*, **56**, 375.
Havinga, E. & de Wael, J. 1940 *Rec. Trav. chim. Pays-Bas*, **59**, 770.
Hughes, T. P. & Whittingham, G. 1942 *Trans. Faraday Soc.* **38**, 9.
Karle, J. 1946 *J. Chem. Phys.* **14**, 297.
Karle, J. & Brockway, O. 1947 *J. Chem. Phys.* **15**, 213.
Menter, J. W. & Sanders, J. V. 1951 *J. Sci. Instrum.* (In the Press).
Murison, C. 1934 *Phil. Mag.* **17**, 201.
Schoon, M. 1938 *Nature*, **142**, 263.
Shooter, K. V. 1949 Private communication.
Tabor, D. 1940 *Nature*, **145**, 308.
Tabor, D. 1941 *Nature*, **147**, 609.
Tanaka, K. 1938 *Mem. Coll. Sci. Kyoto, A*, **21**, 85.
Tanaka, K. 1939 *Mem. Coll. Sci. Kyoto, A*, **22**, 377.
Tanaka, K. 1941 *Mem. Coll. Sci. Kyoto, A*, **23**, 195.
Thiessen, P. A. & Ehrlich, D. 1933 *Z. Phys. Chem. A*, **165**, 453.
Thiessen, P. A. & Klenck, J. 1935 *Z. Phys. Chem. A*, **174**, 335.
Thiessen, P. A. & Spychalski, R. 1931 *Z. Phys. Chem. A*, **156**, 435.
Thiessen, P. A. & Stauff, J. 1936 *Z. Phys. Chem. A*, **176**, 397.
Trillat, J. J. & Motz, H. 1935 *Trans. Faraday Soc.* **31**, 1127.

Structure of thin films of aliphatic esters and alcohols on metals

BY J. V. SANDERS AND D. TABOR

*Research Laboratory on the Physics and Chemistry of Rubbing Solids,
Department of Physical Chemistry, University of Cambridge,
and Tribophysics Division, C.S.I.R.O., Melbourne, Australia*

(Communicated by F. P. Bowden, F.R.S.—Received 9 June 1950)

[Plates 14 and 15]

This paper describes an electron-diffraction study of the structure and orientation of thin films of aliphatic esters and alcohols. The effect of temperature has also been examined by transmission and reflexion methods. Transmission experiments on small crystals of the alcohols show that the orthorhombic structure observed at room temperature gives place to a hexagonal structure at temperatures below the bulk melting-point. These results are analogous to those observed by Müller (1932) with hydrocarbons using X-ray diffraction. When melting occurs there is a marked increase in the lattice parameter, but the packing appears to remain essentially hexagonal. The esters behave similarly.

Reflexion experiments on thin films of alcohols and esters deposited on metal surfaces show that the first molecular layer is generally oriented with the hydrocarbon chains normal to the surface. With alcohols the monolayer loses its orientation at a temperature close to the bulk melting-point of the alcohol, and the disorientation temperature does not depend appreciably on the nature of the metallic substrate. With esters it is difficult to obtain diffraction patterns owing to evaporation, but the evidence again points to a disorientation of the ester at its bulk melting-point. On zinc and cadmium surfaces, however, good orientation may be obtained up to elevated temperatures, and the behaviour is similar to that observed with metallic soaps. The bearing of these results on the nature of the disorientation process is discussed and a comparison made with the boundary lubricating properties of these substances.

Previous work on the temperature variation on the structure of thin layers of long-chain carbon compounds using electron-diffraction techniques has been confined mainly to fatty acids, soaps and paraffins (see for example Tanaka (1938, 1939, 1941), Brummage (1947) and Cowley (1948)). The present work was undertaken as a parallel investigation to that of the fatty acids and soaps already described by Menter & Tabor (1951) and includes a study of octadecyl alcohol ($C_{18}H_{37}OH$, m.p. $56^{\circ}C$), melissyl alcohol ($C_{31}H_{63}OH$, m.p. $67^{\circ}C$) and the esters methyl stearate ($C_{17}H_{35}COOCH_3$, m.p. $38^{\circ}C$), ethyl stearate ($C_{17}H_{35}COOC_2H_5$, m.p. $33^{\circ}C$), and melissyl melissate ($C_{30}H_{61}COOC_{30}H_{61}$, m.p. $81^{\circ}C$).

Two types of experiment have been carried out. In the first, thin crystalline layers of the materials were examined by transmission. In the second, layers deposited on various metal surfaces were examined by reflexion.

TRANSMISSION EXPERIMENTS

Apparatus and preparation of specimens

The experiments were carried out in a camera of the Finch type working in the range 35 to 40 kV. The specimen holder designed for heating the specimens is described elsewhere (Menter & Sanders 1951). The specimen was mounted on a circular

silver grid, $\frac{1}{4}$ in. in diameter. In the heating experiments the temperature of the specimen could be determined to within a few degrees. The structure was examined with the electron beam normal to the grid and successive diffraction patterns observed on a film as the specimen was heated up.

Three methods were used to prepare suitable crystalline films:

(1) A drop of a solution in a suitable solvent was allowed to spread on a clean water surface. Evaporation of the solvent left a thin crystalline film from which a suitable portion could be collected on to a grid. These films collapsed when the crystals melted.

(2) In some cases a small quantity of collodion was added to the solution. The film was formed on a water surface as before, and was often strong enough not to collapse when the crystals melted.

(3) A crystalline film was evaporated directly on to a thin gold film mounted on a specimen grid. In this case the diffraction pattern contained a superimposed gold pattern which could be used as a standard for measurements. Upon heating, melting could be observed without the film collapsing.

Results

Alcohols

At room temperature, the diffraction patterns from specimens of octadecyl alcohol consisted of spots lying in rectangular cross-gratings (figure 1, plate 14). In many cases patterns were obtained from single crystals and crystal mosaics. The spots (h, k, o) were weak or absent where $h + k$ was odd, thus indicating the presence of the chain in the centre of the orthorhombic unit cell. In some single-crystal patterns there were weak reflexions in the forbidden 100 and 010 positions (see Cowley, Rees & Spink 1950). These patterns showed that the alcohol had crystallized in the γ phase (Schoon 1938) with the basal plane parallel to the supporting surface and the carbon chains normal to it, the dimensions of the unit cell being 7.43×5.02 Å.

The other, longer chain-length alcohol, melissyl alcohol, did not form large single crystals so readily. The diffraction patterns comprised spotty rings rather than discrete spots of a single cross-grating. The rings corresponded to the spots of the octadecanol patterns, giving a unit cell size of 7.58×4.96 Å. Again the crystals were orthorhombic, and the basal planes were preferentially parallel to the supporting grid.

Thermal changes

On examination at room temperature it was found that local heating was caused by exposing the same part of the specimen to the electron beam for more than about 20 sec. (figure 1*b*). From the changes in the diffraction pattern it was deduced that films mounted directly on to grids could be heated up to about 40°C in 1 min. by the electron beam. However, when the films were mounted on a good thermal conductor such as a thin gold film, the heating due to the beam was very much less and probably not more than a few degrees. To minimize the effect of this local heating, the specimen was subjected to the beam for only about 5 sec. while a photograph was taken.

As the temperature was slowly raised by heating the specimen holder, the higher-order spots in the diffraction pattern from octadecyl alcohol became gradually less

intense, and the rectangular array changed gradually to a regular hexagonal array. In this state (about 45° C for octadecanol) the unit cell had the dimensions $a = 8.28$, $b = 4.76$ Å, so that $b = \frac{1}{2}a\sqrt{3}$. At this temperature the pattern from the single crystals of octadecyl alcohol comprised a strong 110 ring on which were strong spots in a regular hexagonal array, and a few weak second-order spots. As heating was continued the second order spots faded and the intensity variations on the 110 ring disappeared, so that just below the melting-point the pattern consisted of only a weak sharp 110 ring, the spacing being equal to $a/2 = 4.14$ Å. Melting was observed by a sudden change in this ring to a diffuse halo with a decrease in radius corresponding to a Bragg lattice spacing increase from 4.14 to 4.6 Å.

With melissyl alcohol the transition to hexagonal array was completed at about 55° C, but just below the melting-point a new sharp ring (spacing 4.6 Å) appeared indicating the presence of another form which was stable for a few degrees below the melting point. When the crystal melted the sharp ring changed to a halo of the same spacing (see figure 2*a, b, c* plate 15).

In those cases where the film did not collapse when molten, the diffraction patterns returned again as the crystals cooled and solidified.

Acids

A brief investigation of two fatty acids (stearic and octacosanoic) indicated that their behaviour was similar to that of the alcohols. It was found that the acid and alcohol of similar chain length attained the state of regular hexagonal packing at about the same temperature.

Esters

Thin crystalline films of methyl stearate and ethyl stearate gave very similar diffraction patterns consisting at room temperature of a set of weak arced rings with orders higher than the second very weak. The 110 ring showed strong hexagonal arcing, and there were two short arcs corresponding to the (200) spacing. The appearance of this pattern was strikingly similar to that from octadecanol at a temperature of about 40° C where the spot pattern was not quite a regular hexagonal array. When the specimens were heated, all spots but the 110 faded. The 110 ring became sharp and continuous, the intensity variations having faded at about 30° C. At this temperature the spacing corresponded to a regular hexagonal array. As for the alcohols, melting was observed by the sudden appearance of a halo.

With crystalline films of melissyl melissate the diffraction patterns always comprised sets of continuous rings. The structure appeared to be similar to that of melissyl alcohol, the unit cell dimensions being 7.51×4.95 Å. On heating, the melissyl melissate films showed none of the changes in the diffraction patterns, which were so typical of the other compounds and the 200 and 110 rings remained discrete up to the melting-point. The main results are summarized in table 1.

Discussion

It is seen that these long-chain compounds with a polar group at one end of a paraffin chain crystallize in the γ or vertical phase when prepared by the methods described. The basal plane is parallel to the supporting surface, i.e. parallel to the

water surface or gold film on to which it has been evaporated. It has already been shown by Coumoulos & Rideal (1941) that the esters are stable in this form when they are prepared in continuous films by the Blodgett technique. This form is also

TABLE 1

material	chain length	bulk m.p.	structure at room temperature	hexagonal structure		
				temp. at which completed (° C)	110 spacing when completed (Å)	110 spacing on melting (Å)
alcohols:						
octadecyl	18	56	rect. 7.43 × 5.02	45	4.14	4.6
melissyl	30	67	rect. 7.58 × 4.96	55	4.15	4.6
esters:						
methyl stearate	1 + 17	38	nearly hexagonal	25-30	4.15	4.6
ethyl stearate	2 + 17	33		25-30	4.1	4.6
melissyl melissate	30 + 30	81	rect. 7.51 × 4.95	—	—	—

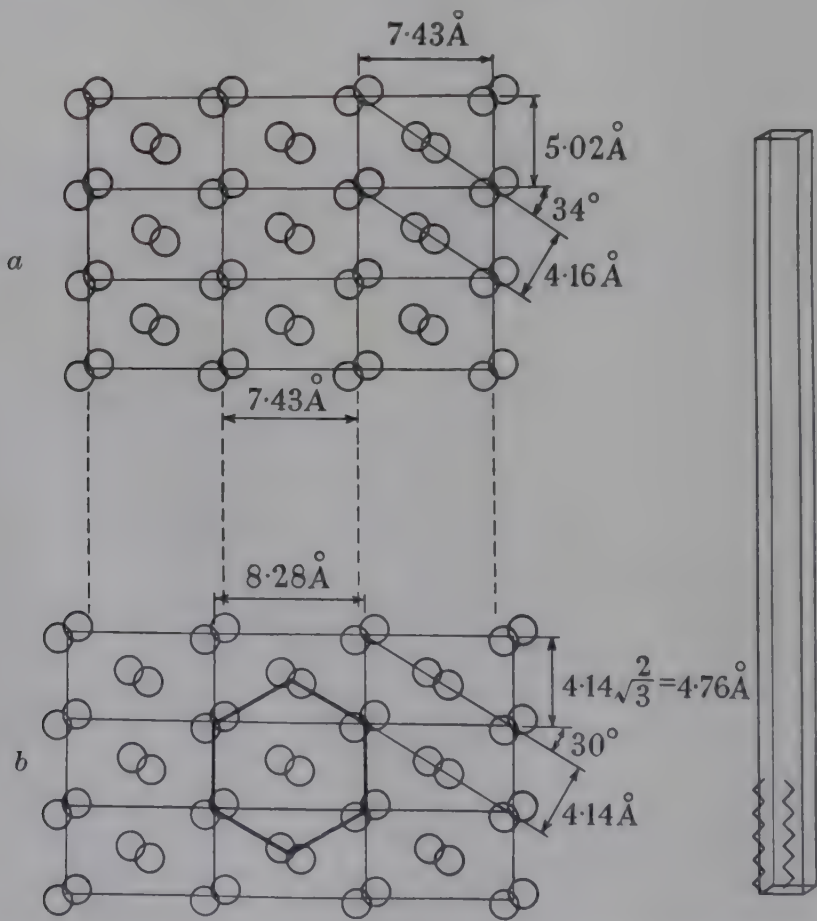


FIGURE 3. Crystal structure of octadecyl alcohol. Left: projection of the unit cells on the basal plane, *a*, at room temperature; *b*, after the change to regular hexagonal array. Right: the orthorhombic unit cell of paraffinic compounds.

known to be stable for the alcohols at room temperature (Wilson & Ott 1934). In many instances large single crystals or mosaics were formed.

When films of all compounds except melissyl melissate were heated, a gradual transition took place in the crystal structure, until at a temperature well below the melting-point the molecules were standing at points which were in a regular hexagonal

array. This regular packing was caused by the a dimension of the unit cell increasing and the b dimension decreasing so that the 110 spacing remained constant (figure 3). Müller (1930, 1932) has shown how some paraffins attain this state of hexagonal packing by unequal expansions of both a and b axes. He found that the paraffin $C_{18}H_{38}$, possessing the same chain length as octadecyl alcohol, melted (at 28°C) before this change was complete. However, it is seen that this transition is completed at 45°C in the alcohol, where the presence of the polar substituent end-group has raised the melting-point above that of the paraffin. The transition is thus fundamentally a property of the paraffin chain. It is a continuous change, and so will not be detected by thermochemical measurements where the methods such as cooling curves detect discontinuous enantiotropic changes only.

The transformation is, however, affected slightly by different substituent end-groups, as is seen by the experiments with the stearates. The structures of these two esters at room temperature were similar to that of octadecyl alcohol at about 35°C . Similarly, the transition to hexagonal array was completed at a lower temperature for the esters (about 30°C) than for the alcohol (about 45°C).

The long-chain ester, melissyl melissate, did not exhibit this transition to a state of hexagonal packing. Müller (1932) has also found that very long-chain paraffins do not undergo this transformation. This suggests that the ester consisting of two equal long hydrocarbon chains joined by an ester group behaves very much like the paraffin of chain length equal to the total length of the ester. On the other hand, melissyl alcohol which occurs as a dimer in the unit cell, behaves similarly to the paraffin of half the dimer length except that its melting-point is increased. It is evident that the chemical link in the melissyl melissate has a much greater influence on the structural properties of the molecule than the dipole attraction in the melissyl alcohol dimer.

REFLEXION EXPERIMENTS

Preparation of specimens

The metal specimens (platinum, copper, cadmium, zinc and mild steel) were in the form of circular disks, $\frac{3}{4}$ in. in diameter and $\frac{1}{8}$ in. thick. The experimental procedure was similar to that described by Menter & Tabor (1951). The layers were deposited by three methods: (1) retraction from the melt, (2) retraction from solution (see Bigelow, Pickett & Zisman 1946), (3) rubbing method.

Experimental results

Alcohols

Layers of octadecyl alcohol (m.p. 56°C), prepared by each of the above methods, were examined on platinum, copper, mild steel, cadmium and zinc. They showed diffraction patterns at room temperature, consisting of straight diffuse bands parallel to the shadow edge. These are typical of monolayers with their hydrocarbon chains perpendicular to the surface. When the specimen was heated, it was found that the monolayer pattern remained substantially unchanged for a considerable rise in temperature. The diffuse bands then curved slightly, and very soon the pattern faded leaving only diffuse scattering and weak rings from the substrate. The

transition temperature was taken as being that temperature at which the second-order diffraction band no longer showed a definite straight portion and could be estimated to within 5°C or less.

It was found that the transition temperature of the monolayer was about 10°C below the bulk melting-point and independent of the method of preparation (see table 2).

TABLE 2. TRANSITION TEMPERATURES ($^{\circ}\text{C}$)

monolayer	m.p. ($^{\circ}\text{C}$)	Pt	Cu	steel	Cd	Zn	preparation
octadecyl alcohol	56	37	40	45	45	45	1, 2, 3
melissyl alcohol	67	70	70	75	75	75	3

Layers of melissyl alcohol (m.p. 67°C) prepared by the rubbing method were examined. It was found that the monolayer disorientation occurred over a range of about 3°C , i.e. more abruptly than for the shorter chain alcohol. The temperature of disorientation on all metals was slightly higher than the bulk melting-point.

Specimens were examined in the diffraction camera after they had cooled and been left to stand for about 24 hr. In all cases a pattern characteristic of a surface film reappeared, indicating that the transition was a true disorientation of the surface film and was not due to evaporation. Monolayer band patterns did not always reappear, however, but for mild steel and copper in particular, arced rings indicative of a polycrystalline layer were observed.

Layers of the alcohols prepared by the rubbing method which had been rubbed only lightly, gave a composite pattern comprising the usual monolayer pattern together with a set of arced rings. These arise from crystals lying above the monolayer, usually with the (110) planes lying parallel to the surface. When such a layer of melissyl alcohol was heated (figure 2, *e, f, g*, plate 15), the arc pattern from the crystals was no longer visible above 40°C ; only the monolayer pattern remained, and this disoriented in the usual way on further heating.

Esters

Ethyl stearate (m.p. 33°C) and methyl stearate (m.p. 38°C) did not form oleophobic monolayers by either method 1 or 2, and so layers were prepared by rubbing down a thicker layer (method 3). Great difficulty was experienced in obtaining any diffraction patterns from layers on platinum and copper. It was found that in the diffraction camera thin layers evaporated from the surface of these metals even at room temperature. Thicker crystalline layers deposited on platinum gave diffraction patterns of slanting bands. When a thick layer of ethyl stearate was heated this band pattern faded gradually and at 34°C a monolayer band appeared. This faded almost immediately and left a diffuse background and a streak, normal to the shadow edge caused by specular reflexion of the electron beam from charged liquid droplets not yet evaporated from the surface.

Good monolayer band patterns were obtained from rubbed layers of both esters on cadmium and zinc. When these specimens were heated no change occurred up to 70°C , after which the first signs of a loss of orientation became apparent. The loss of orientation was slow and the behaviour was similar to that of a monolayer of stearic



a



b



c



d



e



f

FIGURE 1. Transmission patterns from octadecyl alcohol. *a*, pattern predominantly from one crystal; specimen exposed to electron beam for 4 sec.; *b*, same part of specimen exposed to beam for one minute; *c*, polycrystalline layer formed by evaporation of benzene solution on gold support; *d*, same specimen heated to 45° C; *e*, same specimen just below melting-point; *f*, same specimen just above melting-point. *c*, *d*, *e* and *f* were taken with 4 sec. exposure.

(Facing p. 530)

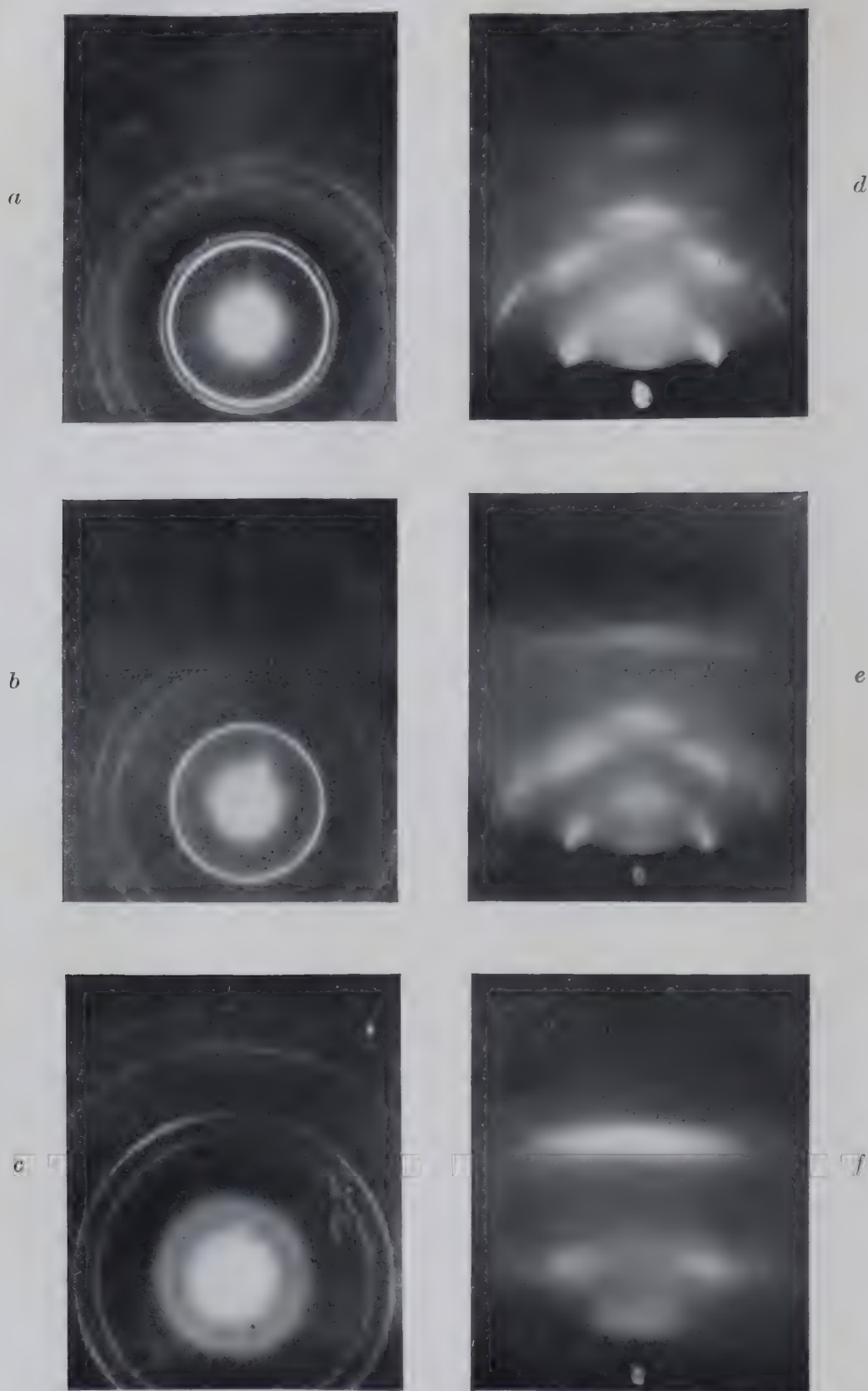


FIGURE 2. *a* (22°C), *b* (57°C), *c* (75°C), transmission patterns of thin film of melissyl alcohol evaporated from benzene solution on to thin gold film. *d* (22°C), *e* (30°C), *f* (52°C), reflexion patterns from melissyl alcohol on mild steel.

acid on these metals (Menter & Tabor 1951). The fading was gradual, and haloes were visible from 100 to about 120° C.

Thick layers of melissyl melissate were prepared by evaporation from a solution on to the metal surfaces. The diffraction patterns comprised bands broken sharply into spots similar to patterns obtained from thin layers of paraffins by Brummage (1947). When these layers were heated the details on the bands faded so that at 60° C the pattern was as from a monolayer. These bands disappeared entirely at 75° C, a few degrees below the bulk melting-point. A rubbed layer of melissyl melissate showed a tilted structure similar to that observed by Menter and Tabor for octocosanoic acid; disorientation again occurred at about 75° C.

Discussion

These experiments show that octadecyl alcohol disorients on all metals at a temperature slightly below the bulk melting-point, and melissyl alcohol very close to the melting-point. It can be concluded from these experiments and from those by Menter & Tabor on fatty acids and soaps that in general a monolayer of a polar compound of the C₁₈ series loses its regular orientation at about 10° C below the bulk melting-point of the compound of which the monolayer is composed. For the longer chains the transition temperature is within a few degrees of the bulk melting-point. Moreover, the nature of the disorientation corresponds to that of the bulk melting, i.e. it is sharp for alcohol and pure acid films, but protracted for soap films.

The method of preparation of the monolayer had no effect on the disorientation temperature for monolayers of (i) alcohols on reactive metals, (ii) alcohols on unreactive metals, (iii) acids on unreactive metals. This shows that there is no inherent difference in the packing of the monolayers prepared by these different methods. With fatty acids on reactive metals, however, the retracted monolayers have a lower disorientation temperature than the rubbed monolayers (Menter & Tabor 1951). This may be due to a better alinement of the molecules in the rubbed monolayer or to more complete soap formation.

The behaviour of methyl and ethyl stearate on reactive metals such as cadmium and zinc is strikingly similar to that of the fatty acids. Disorientation is protracted and occurs at a temperature close to the softening point of the appropriate metallic stearate. The results suggest the formation of the metallic soap, a conclusion which has been directly confirmed by Bowden & Moore (1949) using radioactive tracers. In these tracer experiments the total amount of soap formed after prolonged contact between the metal and the ester corresponded to 8 parts of fatty acid in 10⁶ parts of ester. If the whole of the reaction occurred with fatty acid present as impurity in the ester, this would indicate that the amount of free fatty acid was less than 1 part in 10⁵ (Moore 1949).

These general results are briefly summarized in table 3, the results for acids and soaps being taken from Menter & Tabor.

Thus, taking into account surface reactions which may occur, the thermal behaviour of a monolayer of a long straight-chain carbon compound is very similar to that of the crystals in bulk. This suggests that the structure of a monolayer adsorbed on a metal surface is not widely different from that of the bulk. There is

some evidence (Daniel 1949) that the packing of the chains on the surface is similar to that of the molecules in the crystal, so that the forces between the chains are similar in the two cases. Moreover, in the case of the monolayer the image of the dipole in the surface provides an attractive force equivalent to that between molecules associated in dimers in the solid. Thus the forces of attraction between the molecules are similar. Hence one would expect melting of the solid and disorientation of the monolayer to occur at about the same temperature.

TABLE 3

material	m.p. (° C.)	metal substrate	disorienta- tion temp. (° C)	friction transition temp. (° C)
acid:				
stearic	69	Pt	60	65
stearic	69	Cd	130	100
soap:				
cadmium stearate	110	all metals	100	130
alcohols:				
octadecyl	56	all metals	45	45-55
melissyl	67	all metals	70	—
esters:				
methyl stearate	38	Pt	—	30
ethyl stearate	33	Pt	35	30
ethyl stearate	33	Cd	120	100
melissyl melissate	81	all metals	75	—

The formation of oleophobic monolayers

These results are of interest in connexion with the formation of retracted monolayers. These monolayers may be formed by retraction from a solution or from the melt. In the latter case the excess liquid retracts from the adsorbed monolayer, although the electron diffraction observations show that, at this temperature, the monolayer is strongly disoriented. Thus orientation of the monolayer is not a necessary criterion for the formation of an oleophobic monolayer. Apparently a preferential attachment of the polar end-group is sufficient. It is not surprising, therefore, to find that the temperature at which these oleophobic monolayers are no longer formed by adsorption from the melt (Bigelow, Glass & Zisman 1947) is considerably higher than the disorientation temperature.

Orientation and boundary lubrication

Finally, we may compare the structural properties of adsorbed monolayers with their behaviour as boundary lubricants. When metal surfaces are lubricated with thin films of organic material it is found that the friction does not vary markedly as the temperature is increased, but at a critical temperature there is a sudden increase in the friction and in the amount of surface damage. The frictional breakdown temperatures are given in the fifth column of table 3, and it is seen that there is, in general, close agreement between them and the disorientation temperatures observed by electron diffraction. It would appear from these results that the high breakdown temperatures observed with fatty acids and esters on reactive metals are due

primarily to the formation of a metallic soap at the metal surface. It is also apparent that effective boundary lubrication with long-chain compounds is associated with the presence of a well oriented layer of molecules on the surface.

We wish to thank Dr F. P. Bowden for advice and encouragement, Dr W. Cochrane and Dr J. W. Menter for helpful discussions, the Ministry of Supply (Air) for a Research Grant and C.S.I.R.O. (Australia) for a Studentship to J.V.S.

REFERENCES

- Bigelow, W. C., Glass, E. & Zisman, W. A. 1947 *J. Coll. Sci.* **2**, 563.
 Bigelow, W. C., Pickett, D. L. & Zisman, W. A. 1946 *J. Coll. Sci.* **1**, 513.
 Bowden, F. P. & Moore, A. C. 1949 *Research*, **2**, 585.
 Brummage, K. G. 1947 *Proc. Roy. Soc. A*, **188**, 414; **191**, 243.
 Coumoulos, G. D. & Rideal, E. K. 1941 *Proc. Roy. Soc. A*, **178**, 421.
 Cowley, J. M. 1948 *Trans. Faraday Soc.* **44**, 60.
 Cowley, J. M., Rees, A. L. G. & Spink, J. 1950 (In the Press).
 Daniel, S. G. 1949 Ph.D. Thesis, Cambridge.
 Menter, J. W. & Sanders, J. V. 1951 *J. Sci. Instrum.* (In the Press).
 Menter, J. W. & Tabor, D. 1951 *Proc. Roy. Soc. A*, **204**, 514.
 Moore, A. C. 1949 Ph.D. Dissertation, Cambridge.
 Müller, A. 1930 *Proc. Roy. Soc. A*, **127**, 417.
 Müller, A. 1932 *Proc. Roy. Soc. A*, **138**, 521.
 Schoon, T. 1938 *Z. phys. Chem. B* **39**, 385.
 Tanaka, K. 1938 *Mem. Coll. Sci. Kyoto, A*, **21**, 85.
 Tanaka, K. 1939 *Mem. Coll. Sci. Kyoto, A*, **22**, 377.
 Tanaka, K. 1941 *Mem. Coll. Sci. Kyoto, A*, **23**, 195.
 Wilson, D. A. & Ott, E. 1934 *J. Chem. Phys.* **2**, 231.

Light distribution near focus in an error-free diffraction image

BY E. WOLF, *The Observatory, University of Cambridge*

(Communicated by R. O. Redman, F.R.S.—Received 16 June 1950)

Expressions are derived for the fraction of the total illumination present within certain regions in receiving planes near focus of spherical waves issuing from a circular aperture. The derivation is based on Lommel's treatment of Fresnel diffraction, and the solution takes the form of rapidly convergent series involving Bessel functions. Numerical results are illustrated by contour lines. The distribution of the illumination in a number of selected planes near focus is studied in greater detail. A comparison with the predictions of geometrical optics is also made. As a special case the fraction of the total illumination present in the geometrical shadow is discussed. The results provide a mathematical basis for a discussion of the imaging properties in optical systems where, as for example in a well-corrected refracting telescope objective, the chromatic variation of focus is the only appreciable aberration.

1. INTRODUCTION

Although a good deal of attention has been paid in recent years to problems concerning the effect of diffraction on monochromatic image formation in the presence of aberrations, a number of important questions dealing with the diffraction of

perfectly spherical waves have so far not been treated. Consider, for example, the diffraction of such waves at a circular aperture. With the usual approximations the intensity distribution in the geometrical focal plane is given by the well-known formula due to Airy (1835). The distribution of intensity in other planes was discussed by Lommel (1885) in a classical memoir and more recently by Nijboer (1942) and Zernike & Nijboer (1949). In most practical applications, however, it is desirable to know not only the intensity at points near focus of the waves but also the total illumination in the various rings of the diffraction pattern. In the Fraunhofer case, where the intensity is given by Airy's expression, the fraction L of the total illumination† which is received within a circle of radius r and centred on the axis is given by the formula due to Rayleigh (1881),

$$L = 1 - J_0^2(v) - J_1^2(v). \quad (1.1)$$

Here $v = \frac{2\pi R}{\lambda f} r$, R being the aperture radius, f the radius of the spherical wave filling the opening and λ the wave-length. The J 's denote Bessel functions of the first kind. The literature contains no corresponding result for the more general Fresnel case. In this case the total illumination could of course be calculated by numerical integrations from Lommel's formulae or from those of Nijboer and Zernike. But such a procedure would involve a prohibitive amount of numerical work because of the form of the expressions and the complicated behaviour of the intensity distribution (see figure 2). In the present paper formulae are derived which permit direct calculation of the total illumination.

Starting from Lommel's expressions we obtain a solution which takes the form of rapidly convergent series involving Bessel functions. The solution is evaluated near focus, and the results are displayed in figure 3*a* in the form of contour lines. A comparison with results predicted by geometrical optics is also made. It is clearly seen that although in the immediate neighbourhood of the focus the illumination curves predicted by geometrical optics do not resemble those of the physical solution, the two solutions approach each other more and more closely as the distance from the focus becomes large compared with the wave-length. The behaviour of L in several planes near the focus is examined in greater detail.

In the special case when the circle for which the fraction of the total illumination is determined coincides with the geometrical confusion disk, the series can be summed. The solution then reduces to the simple form

$$L = 1 - J_0(v) \cos v - J_1(v) \sin v, \quad (1.2)$$

the expression $J_0(v) \cos v + J_1(v) \sin v$ giving the fraction of light within the geometrical shadow.

The results derived in the present paper supplement those of Lommel, Nijboer and Zernike and, with them, provide the mathematical basis for a quantitative discussion

† Following Rayleigh's usage we define the total illumination in a region D as

$$\iint_D I(x, y) \, dx \, dy,$$

where $I(x, y)$ is the intensity at a typical point (x, y) of D .

of the imaging properties in optical systems where chromatic variation of focus is the only appreciable aberration. For example, they make it practicable to calculate the light scattered by diffraction from stellar images formed by refracting telescope objectives, taking into consideration the presence of the secondary spectrum.

2. EXPRESSIONS FOR THE TOTAL ILLUMINATION

In what follows, u and v are real variables, n and m are non-negative integers, J_n denotes a Bessel function of the first kind and U_n and V_n are Lommel functions defined by†

$$\begin{aligned} U_n(u, v) &= \sum_{s=0}^{\infty} (-1)^s \left(\frac{u}{v}\right)^{n+2s} J_{n+2s}(v), \\ V_n(u, v) &= \sum_{s=0}^{\infty} (-1)^s \left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v). \end{aligned} \quad (2.1)$$

Further, Y_n and W_n denote the related functions

$$Y_n(u, v) = \sum_{s=0}^{\infty} (-1)^s (n+2s) \left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v), \quad (2.2)$$

$$W_n(u, v) = \sum_{s=0}^{\infty} (-1)^s (s+1) \left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v). \quad (2.3)$$

We also define the polynomials

$$P_{n,2m}(v) = \sum_{s=0}^{2m} (-1)^s J_{n+s}(v) J_{n+2m-s}(v). \quad (2.4)$$

Finally, we shall find it convenient to set

$$Q_{2m}(v) = P_{0,2m}(v) + P_{1,2m}(v). \quad (2.5)$$

(a) We shall be concerned with the light distribution in a receiving plane near focus O , of convergent spherical waves which issue from a circular aperture. Let R be the radius of the opening, C the point of intersection between the spherical wave and the axis joining O to the centre of the aperture, and let $CO = f$. Further, let O' denote the point of intersection of the axis with the receiving plane, both this plane and the plane of the aperture being assumed to be perpendicular to the axis (figure 1).

To describe the effect of the light waves at a point P near the focus, we introduce polar co-ordinates (r, θ) in the receiving plane with origin at O' and set

$$u = \frac{2\pi R^2}{\lambda f^2} \Delta f, \quad v = \frac{2\pi R}{\lambda f} r, \quad (2.6)$$

† Two slightly different definitions of the V_n function appear in the literature. That given by (2.1) is used throughout this paper. The Y_n function has recently been introduced by Hopkins (1949) in a paper concerning waves of non-uniform amplitudes. These functions are related by the equation

$$Y_n(u, v) = \frac{1}{2} \left[\frac{v^2}{u} V_{n-1}(u, v) + u V_{n+1}(u, v) \right]$$

which is implicit in Lommel's memoir and which follows from (2.2) by the application of the identity

$$(n+2s) J_{n+2s}(v) = \frac{1}{2} v [J_{n+2s-1}(v) + J_{n+2s+1}(v)].$$

where $\Delta f = OO'$ (positive in figure 1) denotes the distance between the receiving plane and the geometrical focal plane $u = 0$; it is assumed throughout this paper that $\Delta f/f$, r/f and R/f are small compared to unity. In this notation the first zero of intensity on the axis is given by $u = 4\pi$, whilst the first dark ring in the geometrical focal plane (i.e. the Airy ring) is given by $v = 1.22\pi$. The outline of the geometrical shadow is given by $u = \pm v$.

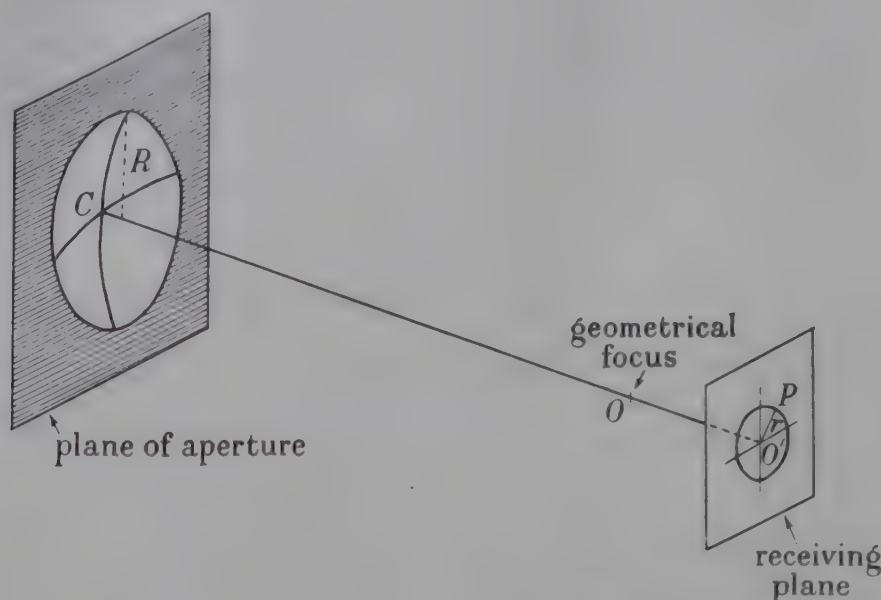


FIGURE 1. $CO = f$; $OO' = \Delta f$; $u = \frac{2\pi R^2}{\lambda f^2} \Delta f$; $v = \frac{2\pi R}{\lambda f} r$.

The physical significance of the parameters u and v is as follows: If we test in an interferometer the wave in the opening against a spherical wave centred at P , then $u/4\pi$ is the number of fringes of defocusing whilst v/π is the number of fringes of lateral displacement of P from the axis.

Let $\frac{A}{f} \sin \frac{2\pi}{\lambda} (ct - f)$ be the wave disturbance in the opening. Here A is a constant while c and t denote the velocity of light and the time respectively. The intensity $I(u, v)$ at the point P is then given, in suitable units, by the following formula due to Lommel† (1885):

$$I(u, v) = \frac{4\gamma}{u^2} [U_1^2(u, v) + U_2^2(u, v)], \quad (2.7)$$

where

$$\gamma = \left(\frac{A\pi R^2}{\lambda f^2} \right)^2,$$

and the U 's are Lommel functions defined by (2.1). When $|u/v| > 1$ the U -series converge too slowly to be useful for computation, and for this case Lommel expressed the solution by means of the related V functions, in the equivalent form

$$I(u, v) = \frac{4\gamma}{u^2} \left[1 + V_0^2(u, v) + V_1^2(u, v) - 2V_0(u, v) \cos \frac{1}{2} \left(u + \frac{v^2}{u} \right) - 2V_1(u, v) \sin \frac{1}{2} \left(u + \frac{v^2}{u} \right) \right]. \quad (2.8)$$

† The solutions (2.7) and (2.8) were also obtained by Struve (1886) by analysis similar to that of Lommel.

Recently Nijboer (1942) and Zernike & Nijboer (1949) derived in the course of a more general investigation another expression for the intensity and gave a figure showing the isophotes near focus (figure 2).

Starting from Lommel's equations we shall now derive expressions for the total illumination received within the area in the receiving plane which is bounded by the circle centred on O' and of radius r_0 . We define

$$L(u, v_0) = \frac{1}{B} \int_0^{r_0} \int_0^{2\pi} I(u, v) r dr d\theta, \quad (2.9)$$



FIGURE 2. Isophotes [contour lines of $I(u, v)$] near focus in a meridional plane. The straight lines indicate the boundary of the geometrical shadow. The numbers give intensity as percentage of the intensity at focus. Axial minima and maxima are indicated by short strokes, others by small circles. The figure covers approximately the range $-35 \leq u \leq 35$, $0 \leq v \leq 15$. (After Zernike & Nijboer 1949.)

where v_0 denotes the value of the v parameter for points on the boundary of the circle and

$$B = \left(\frac{A}{f}\right)^2 \pi R^2.$$

To the order of accuracy here in question, $L(u, v_0)$ may be regarded as giving the fraction of the total energy radiated through the aperture which reaches the circle centred at O' . On integrating (2.9) with respect to θ , it follows that

$$L(u, v_0) = \frac{1}{2B} \int_0^{v_0} I(u, v) v dv. \quad (2.10)$$

The boundary of the circle over which the illumination is to be determined lies inside or outside the geometrical confusion disk according as $|v_0/u| \leq 1$, and it coincides with its boundary when $|v_0/u| = 1$. We shall now examine the three cases separately. Since, within the range in which (2.7) and (2.8) are valid, $I(u, v)$ and consequently $L(u, v)$ are even functions of u , we can without loss of generality assume u to be positive. The parameter v is positive by definition.

(1) *The case $v_0/u < 1$*

In this case we have from (2.8) and (2.10)

$$L(u, v_0) = \frac{2}{u^2} [\frac{1}{2}v_0^2 + L_1(u, v_0) - 2L_2(u, v_0) - 2L_3(u, v_0)], \quad (2.11)$$

where

$$\left. \begin{aligned} L_1(u, v_0) &= \int_0^{v_0} v [V_0^2(u, v) + V_1^2(u, v)] dv, \\ L_2(u, v_0) &= \int_0^{v_0} v V_0(u, v) \cos \frac{1}{2} \left(u + \frac{v^2}{u} \right) dv, \\ L_3(u, v_0) &= \int_0^{v_0} v V_1(u, v) \sin \frac{1}{2} \left(u + \frac{v^2}{u} \right) dv. \end{aligned} \right\} \quad (2.12)$$

To evaluate L_1 we apply lemma 2† and obtain

$$L_1(u, v_0) = \int_0^{v_0} \sum_{s=0}^{\infty} (-1)^s \left(\frac{v}{u} \right)^{2s} P_{0, 2s}(v) v dv. \quad (2.13)$$

We find from lemma 4 that in the range of integration the series under the integral sign is dominated by the series

$$\sum_{s=0}^{\infty} \frac{2s+1}{s!} \left(\frac{v_0^2}{2u} \right)^{2s} \exp \left(\frac{1}{2} v_0^2 \right),$$

which converges for all values of v_0 . Hence (2.13) may be integrated term by term. By lemma 6 and from the definition of Q it then follows that

$$L_1(u, v_0) = \frac{v_0^2}{2} \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} \left(\frac{v_0}{u} \right)^{2s} Q_{2s}(v_0). \quad (2.14)$$

To evaluate L_2 , we write it in the form

$$L_2(u, v_0) = \cos \frac{1}{2} u \int_0^{v_0} v V_0(u, v) \cos \frac{v^2}{2u} dv - \sin \frac{1}{2} u \int_0^{v_0} v V_0(u, v) \sin \frac{v^2}{2u} dv. \quad (2.15)$$

We have from lemma 9, for $n = 0$, on equating real and imaginary parts that

$$\left. \begin{aligned} \int_0^{v_0} v V_0(u, v) \cos \frac{v^2}{2u} dv &= u \left[W_1(u, v_0) \cos \frac{v_0^2}{2u} + W_2(u, v_0) \sin \frac{v_0^2}{2u} \right], \\ \int_0^{v_0} v V_0(u, v) \sin \frac{v^2}{2u} dv &= u \left[W_1(u, v_0) \sin \frac{v_0^2}{2u} - W_2(u, v_0) \cos \frac{v_0^2}{2u} \right]. \end{aligned} \right\} \quad (2.16)$$

(2.15) and (2.16) give

$$L_2(u, v_0) = u \left[W_1(u, v_0) \cos \frac{1}{2} \left(u + \frac{v_0^2}{u} \right) + W_2(u, v_0) \sin \frac{1}{2} \left(u + \frac{v_0^2}{u} \right) \right]. \quad (2.17)$$

In a similar way we find

$$L_3(u, v_0) = u \left[W_2(u, v_0) \sin \frac{1}{2} \left(u + \frac{v_0^2}{u} \right) - W_3(u, v_0) \cos \frac{1}{2} \left(u + \frac{v_0^2}{u} \right) \right]. \quad (2.18)$$

† Lemmas quoted without other reference will be found in the appendix.

Substituting for L_1 , L_2 and L_3 into (2.11) we obtain

$$L(u, v_0) = \left(\frac{v_0}{u}\right)^2 \left[1 + \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} \left(\frac{v_0}{u}\right)^{2s} Q_{2s}(v) \right] - \frac{4}{u} \left\{ [W_1(u, v_0) - W_3(u, v_0)] \cos \frac{1}{2} \left(u + \frac{v_0^2}{u}\right) + 2W_2(u, v_0) \sin \frac{1}{2} \left(u + \frac{v_0^2}{u}\right) \right\}. \quad (2.19)$$

But $W_1(u, v_0) - W_3(u, v_0)$

$$\begin{aligned} &= \sum_{s=0}^{\infty} (-1)^s (s+1) \left(\frac{v_0}{u}\right)^{2s+1} J_{2s+1}(v_0) - \sum_{s=0}^{\infty} (-1)^s (s+1) \left(\frac{v_0}{u}\right)^{2s+3} J_{2s+3}(v_0) \\ &= \left(\frac{v_0}{u}\right) J_1(v_0) + \sum_{s=1}^{\infty} (-1)^s (s+1) \left(\frac{v_0}{u}\right)^{2s+1} J_{2s+1}(v_0) - \sum_{s=1}^{\infty} (-1)^{s-1} s \left(\frac{v_0}{u}\right)^{2s+1} J_{2s+1}(v_0) \\ &= \sum_{s=0}^{\infty} (-1)^s (2s+1) \left(\frac{v_0}{u}\right)^{2s+1} J_{2s+1}(v_0) \\ &= Y_1(u, v_0). \end{aligned} \quad (2.20)$$

Also

$$\begin{aligned} W_2(u, v_0) &= \sum_{s=0}^{\infty} (-1)^s (s+1) \left(\frac{v_0}{u}\right)^{2s+2} J_{2s+2}(v_0) \\ &= \frac{1}{2} Y_2(u, v_0). \end{aligned} \quad (2.21)$$

Substituting from (2.20) and (2.21) into (2.19), we finally obtain for the case $v_0/u \leq 1$.

$$L(u, v_0) = \left(\frac{v_0}{u}\right)^2 \left[1 + \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} \left(\frac{v_0}{u}\right)^{2s} Q_{2s}(v_0) \right] - \frac{4}{u} \left[Y_1(u, v_0) \cos \frac{1}{2} \left(u + \frac{v_0^2}{u}\right) + Y_2(u, v_0) \sin \frac{1}{2} \left(u + \frac{v_0^2}{u}\right) \right]. \quad (2.22)$$

(2) The case $v_0/u = 1$

In this case the boundary of the circle coincides with the edge of the geometrical shadow. The expression for L can then be obtained by putting $v_0 = u$ in (2.22). We then find

$$L(u, u) = 1 + \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) - \frac{4}{u} [Y_1(u, u) \cos u + Y_2(u, u) \sin u]. \quad (2.23)$$

Now

$$Y_n(u, u) = \sum_{s=0}^{\infty} (-1)^s (n+2s) J_{n+2s}(u), \quad (2.24)$$

and it has been shown by Lommel (1868, (9), p. 40) that for $n \geq 1$ the sum of the series on the right-hand side of (2.24) is

$$\frac{u}{2} J_{n-1}(u).$$

$$\text{Hence for } n \geq 1, \quad Y_n(u, u) = \frac{u}{2} J_{n-1}(u). \quad (2.25)$$

We also have, by lemma 7, that

$$\sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) = J_0(u) \cos u + J_1(u) \sin u. \quad (2.26)$$

Substituting from the last two equations into (2.23) we find that

$$L(u, u) = 1 - J_0(u) \cos u - J_1(u) \sin u. \quad (2.27)$$

(3) *The case $v_0/u > 1$*

In this case we find it convenient to split the range of integration into the two intervals $0 \leq v \leq u$ and $u < v \leq v_0$. We then have from (2.7), (2.10) and (2.27) that

$$L(u, v_0) = 1 - J_0(u) \cos u - J_1(u) \sin u + \frac{2}{u^2} \int_0^{v_0} v [U_1^2(u, v) + U_2^2(u, v)] dv. \tag{2.28}$$

From lemma 1,

$$\int_u^{v_0} v [U_1^2(u, v) + U_2^2(u, v)] dv = \int_0^{v_0} \sum_{s=0}^\infty (-1)^s \left(\frac{u}{v}\right)^{2s+2} P_{1,2s}(v) v dv.$$

With the help of lemma 4 it can be shown that the series under the integral sign is uniformly convergent in the range of integration and can therefore be integrated term by term. We then find

$$\begin{aligned} \frac{2}{u^2} \int_0^{v_0} v [U_1^2(u, v) + U_2^2(u, v)] dv &= 2 \sum_{s=0}^\infty (-1)^s u^{2s} \int_u^{v_0} v^{-2s-1} P_{1,2s}(v) dv \\ &= \sum_{s=0}^\infty \frac{(-1)^s}{2s+1} u^{2s} \left[\frac{Q_{2s}(u)}{u^{2s}} - \frac{Q_{2s}(v_0)}{v_0^{2s}} \right] \quad (\text{by lemma 5}) \\ &= \sum_{s=0}^\infty \frac{(-1)^s}{2s+1} Q_{2s}(u) - \sum_{s=0}^\infty \frac{(-1)^s}{2s+1} \left(\frac{u}{v_0}\right)^{2s} Q_{2s}(v_0) \\ &= J_0(u) \cos u + J_1(u) \sin u - \sum_{s=1}^\infty \frac{(-1)^s}{2s+1} \left(\frac{u}{v_0}\right)^{2s} Q_{2s}(v_0) \\ &\hspace{15em} (\text{by lemma 7}). \end{aligned} \tag{2.29}$$

Substituting from (2.29) into (2.28), we finally obtain for the case $v_0/u > 1$

$$L(u, v_0) = 1 - \sum_{s=0}^\infty \frac{(-1)^s}{2s+1} \left(\frac{u}{v_0}\right)^{2s} Q_{2s}(v_0). \tag{2.30}$$

(b) Before we discuss the general solution (2.22), (2.27) and (2.30) we shall derive expressions for the total illumination according to geometrical optics.

The rays of light which proceed from the region of the opening between circles of radii h and $h + dh$ centred on the axis cut the receiving plane in points of the annulus formed by the circles of radii r and $r + dr$ centred at O' , where, to first order,

$$\frac{r}{\Delta f} = \frac{h}{f} \quad (h \leq R). \tag{2.31}$$

The intensity I^* at P , according to geometrical optics, is then given by

$$\begin{aligned} I^*(u, v) &= \frac{2\pi h dh}{2\pi r dr} \frac{A^2}{f^2} \quad \text{when } r < \frac{R}{f} |\Delta f|, \\ &= 0 \quad \text{when } r > \frac{R}{f} |\Delta f|. \end{aligned} \tag{2.32}$$

From (2.31) and (2.32) it then follows on eliminating h , that

$$\begin{aligned} I^*(u, v) &= \left(\frac{A}{\Delta f}\right)^2 = \frac{4\gamma}{u^2} \quad \text{when } \left|\frac{v}{u}\right| < 1, \\ &= 0 \quad \text{when } \left|\frac{v}{u}\right| > 1. \end{aligned} \tag{2.33}$$

The series present in the expressions for L were found to converge very rapidly. This is due to the fact that since

$$|J_r(v)| \leq 1, \tag{3.2}$$

$$|Q_{2m}(v)| \leq \sum_{s=0}^{2m} |J_s(v) J_{2m-s}(v)| + \sum_{s=0}^{2m} |J_{s+1}(v) J_{2m+1-s}(v)| \leq 2(2m+1), \tag{3.3}$$

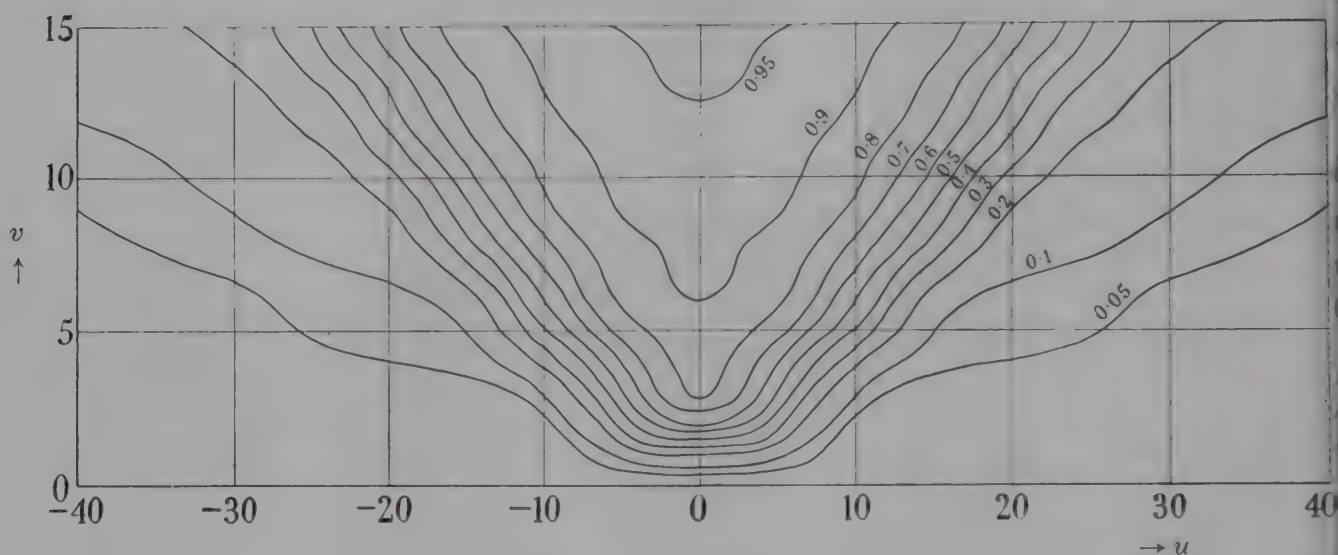
and consequently in the required range, the series

$$\sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} \left(\frac{A}{B}\right)^{2s+1} Q_{2s}(v)$$

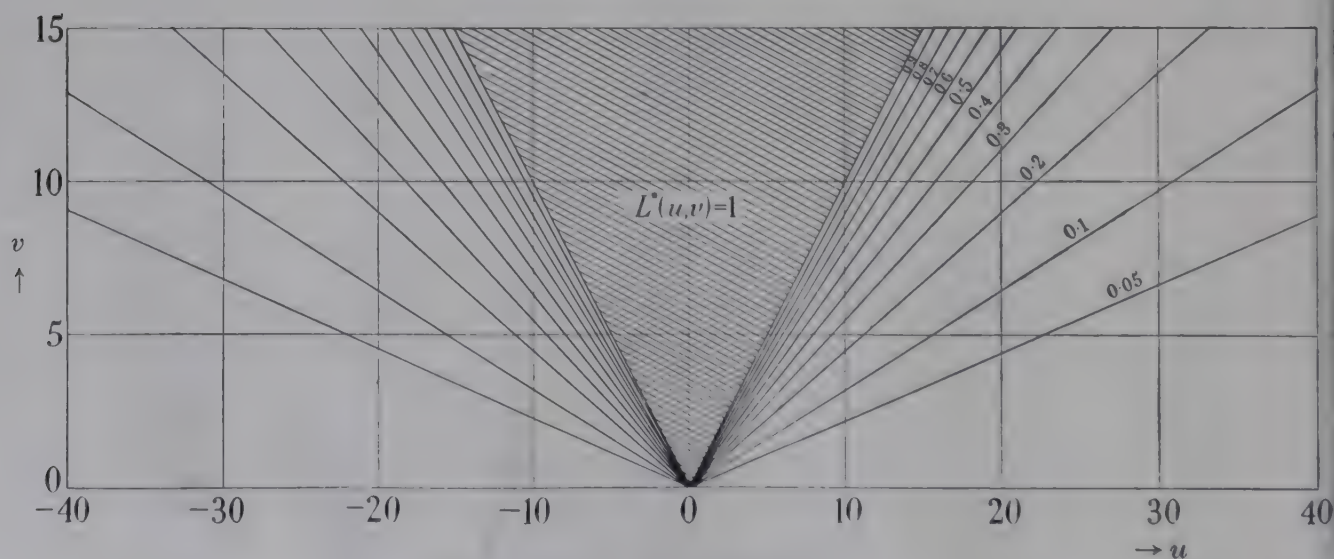
are more rapidly convergent than the geometrical series

$$2 \sum_{s=0}^{\infty} (-1)^s t^{2s+1},$$

where $0 \leq t < 1$. Similar considerations apply to the functions Y_1 and Y_2 in (2.22).



(a). Contour lines of $L(u, v)$



(b). Contour lines of $L^*(u, v)$

FIGURE 3. Contour lines for the fraction of total illumination in circles centred on axis in receiving planes near focus, according to physical optics and according to geometrical optics. Throughout the shaded region $L^*(u, v) = 1$.

From the calculated values the contour lines of L were constructed for the range $-40 \leq u \leq 40$, $0 \leq v \leq 15$ by graphical interpolation. These are displayed in figure 3a; the corresponding curves predicted by geometrical optics are shown in figure 3b for comparison. It is seen that in the immediate neighbourhood of the focus geometrical optics does not give a good approximation to the physical solution. With increasing departure from the focus the two solutions approach each other more and more closely, the level curves of L being in good qualitative agreement with those of L^* already at distances a few fringes away from the focus.

The behaviour of L and L^* in the planes $u = 0, 2\pi, 4\pi, 6\pi, 8\pi, 10\pi$ and 12π corresponding to defocusing of $0, \frac{1}{2}, 1, 1\frac{1}{2}, 2, 2\frac{1}{2}$ and 3 fringes respectively is shown in figures 4a and 4b. In the special case when $u = 0$, the receiving plane coincides with the geometrical focal plane and L is then given by (2.30) with $u = 0$, viz. by

$$L(0, v) = 1 - Q_0(v) = 1 - J_0^2(v) - J_1^2(v) \quad (3.4)$$

in agreement with Rayleigh's formula (1.1).

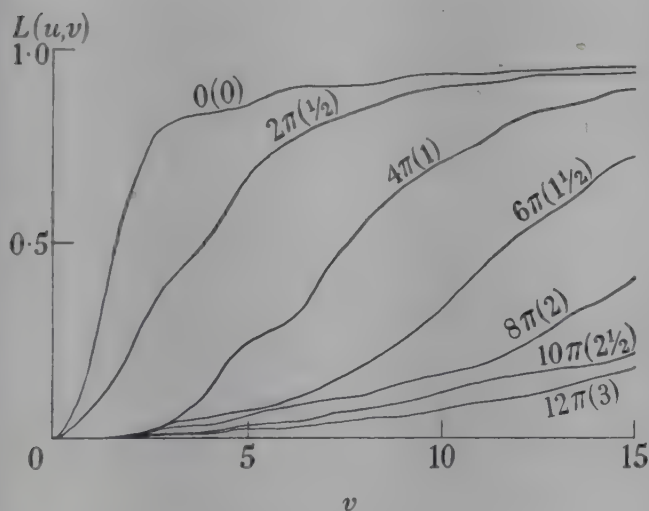


FIGURE 4a. $L(u, v)$

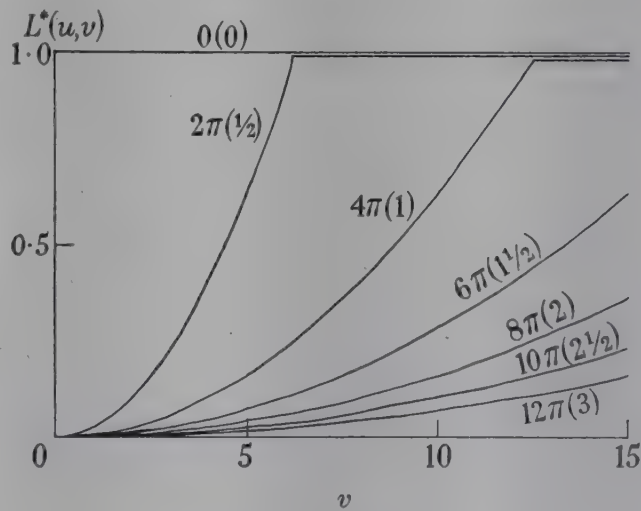


FIGURE 4b. $L^*(u, v)$

The ratio of total illumination in circles centred on axis in receiving planes near focus, according to physical optics (figure 4a) and according to geometrical optics (figure 4b). The numbers along the curves are the values of u while the numbers in brackets denote defocusing in fringes. The horizontal parts of the curves $u = 2\pi$ and $u = 4\pi$ in figure 4b correspond to $L^* = 1$.

It is of interest to examine what fraction $\epsilon(u)$ of the total energy is present within the geometrical shadow. From (2.27) it follows that to a sufficient approximation

$$\epsilon(u) = 1 - L(u, u) = J_0(u) \cos u + J_1(u) \sin u, \quad (3.5)$$

where, as before, u specifies the position of the receiving plane. This function is displayed in figure 5. It is not a strictly decreasing function but it has maxima (apart from $u = 0$) when $J_1(u) = 0$ and minima when $\sin u = 0$ ($u \neq 0$).

We note that for the particular case of diffraction of spherical waves at a circular aperture our solution illustrates the well-known theorem that geometrical optics may be regarded as the limiting case of physical optics as $\lambda \rightarrow 0$. For if we keep r and

Δf fixed and let $\lambda \rightarrow 0$, then $u \rightarrow \infty$ and $v \rightarrow \infty$, but u/v remains finite and it can be verified that in each of the equations (2.22), (2.27) and (2.30) all the terms except the first tend to zero. Consequently as $\lambda \rightarrow 0$,

$$\begin{aligned} L(u, v_0) &\rightarrow \left(\frac{v_0}{u}\right)^2 \text{ when } \left|\frac{v_0}{u}\right| < 1, \\ &\rightarrow 1 \quad \text{when } \left|\frac{v_0}{u}\right| \geq 1, \end{aligned}$$

in agreement with the solution (2.34) of geometrical optics.

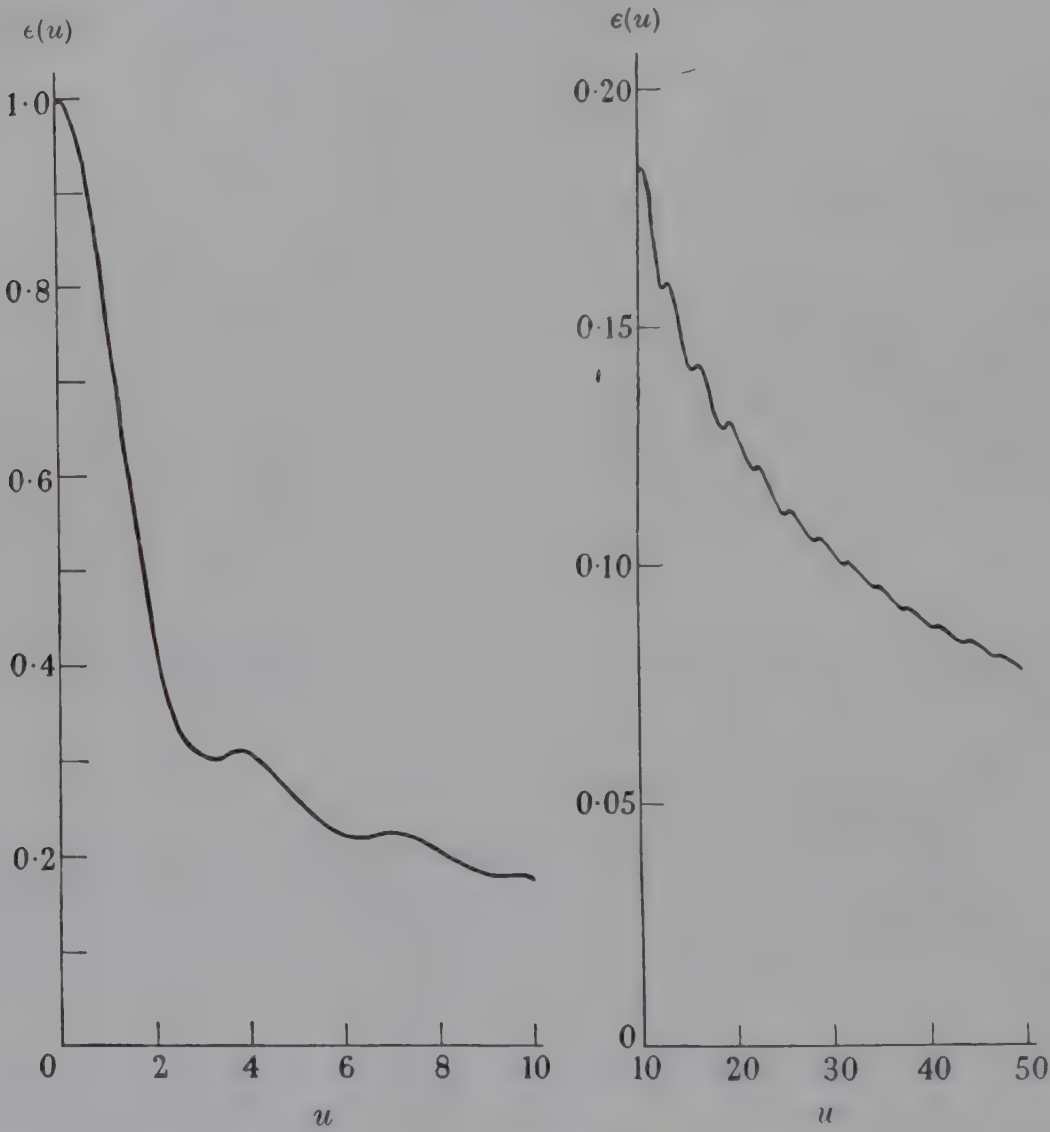


FIGURE 5. The fraction $\epsilon(u)$ of the total energy within the geometrical shadow.

APPENDIX. SOME MATHEMATICAL LEMMAS

The functions U , V , Y , W , P and Q which occur in the lemmas we now derive are defined by equations (2.1) to (2.5)

LEMMA 1.
$$U_n^2(u, v) + U_{n+1}^2(u, v) = \sum_{s=0}^{\infty} (-1)^s \left(\frac{u}{v}\right)^{2(n+s)} P_{n, 2s}(v).$$

Proof. Since the U series are absolutely convergent for all u and v , a series for $U_n^2 + U_{n+1}^2$ can be obtained by performing the various operations term by term.

$$\text{Let } \alpha_s = (-1)^s \left(\frac{u}{v}\right)^{n+2s} J_{n+2s}(v), \quad \beta_s = (-1)^s \left(\frac{u}{v}\right)^{n+1+2s} J_{n+1+2s}(v).$$

$$\text{Then } U_n^2 = \sum_{s=0}^{\infty} a_s, \quad U_{n+1}^2 = \sum_{s=0}^{\infty} b_s,$$

$$\text{where } a_s = \sum_{i=0}^s \alpha_i \alpha_{s-i} = (-1)^s \left(\frac{u}{v}\right)^{2(n+s)} \sum_{i=0}^s J_{n+2i}(v) J_{n+2s-2i}(v),$$

$$b_s = \sum_{i=0}^s \beta_i \beta_{s-i} = (-1)^s \left(\frac{u}{v}\right)^{2(n+1+s)} \sum_{i=0}^s J_{n+1+2i}(v) J_{n+1+2s-2i}(v).$$

$$\text{Hence } U_n^2 + U_{n+1}^2 = \sum_{s=0}^{\infty} a_s + \sum_{s=0}^{\infty} b_s = a_0 + \sum_{s=1}^{\infty} (a_s + b_{s-1}). \quad (\text{A } 1)$$

$$\text{Now } a_0 = \left(\frac{u}{v}\right)^{2n} J_n^2(v) = \left(\frac{u}{v}\right)^{2n} P_{n,0}(v),$$

$$a_s + b_{s-1} = (-1)^s \left(\frac{u}{v}\right)^{2(n+s)} \sum_{i=0}^{2s} (-1)^i J_{n+i}(v) J_{n+2s-i}(v) = (-1)^s \left(\frac{u}{v}\right)^{2(n+s)} P_{n,2s}(v) \quad \text{for } s \geq 1.$$

On substituting from the last two equations into (A 1) lemma 1 follows.

$$\text{LEMMA 2. } V_n^2(u, v) + V_{n+1}^2(u, v) = \sum_{s=0}^{\infty} (-1)^s \left(\frac{v}{u}\right)^{2(n+s)} P_{n,2s}(v).$$

The proof of this lemma is similar to that of lemma 1, v/u and V_n now replacing u/v and U_n respectively.

$$\text{LEMMA 3. } \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v \cos \theta) \frac{\cos(2m+1)\theta}{\cos \theta} d\theta = P_{n,2m}(v).$$

Proof. By definition,

$$P_{n,2m}(v) = \sum_{s=0}^{2m} (-1)^s J_{n+s}(v) J_{n+2m-s}(v).$$

From Watson (1922, 5.43 (1), p. 150),

$$J_{n+s}(v) J_{n+2m-s}(v) = \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v \cos \theta) \cos 2(m-s)\theta d\theta.$$

$$\begin{aligned} \text{Hence } P_{n,2m}(v) &= \sum_{s=0}^{2m} (-1)^s \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v \cos \theta) \cos 2(m-s)\theta d\theta \\ &= \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} J_{2n+2m}(2v \cos \theta) \sigma_m(\theta) d\theta, \end{aligned} \quad (\text{A } 2)$$

$$\text{where } \sigma_m(\theta) = \sum_{s=0}^{2m} (-1)^s \cos 2(m-s)\theta$$

$$= \frac{\cos(2m+1)\theta}{\cos \theta}. \quad (\text{A } 3)$$

Lemma 3 now follows from (A 2) and (A 3).

LEMMA 4. $|P_{n,2m}(v)| \leq \frac{2m+1}{(n+m)!} (\frac{1}{2}v)^{2(n+m)} \exp(\frac{1}{2}v^2).$

Proof. Since $P_{n,2m}(v) = \sum_{s=0}^{2m} (-1)^s J_{n+s}(v) J_{n+2m-s}(v),$

it follows that $|P_{n,2m}(v)| \leq \sum_{s=0}^{2m} |J_{n+s}(v) J_{n+2m-s}(v)|.$

By Watson (1922, (4), p. 16) we have, for $r \geq 0,$

$$|J_r(v)| \leq \frac{(\frac{1}{2}v)^r}{r!} \exp(\frac{1}{4}v^2), \quad (\text{A } 4)$$

so that
$$\begin{aligned} |P_{n,2m}(v)| &\leq \sum_{s=0}^{2m} \frac{(\frac{1}{2}v)^{n+s}}{(n+s)!} \exp(\frac{1}{4}v^2) \frac{(\frac{1}{2}v)^{n+2m-s}}{(n+2m-s)!} \exp(\frac{1}{4}v^2) \\ &= \left(\frac{v}{2}\right)^{2(n+m)} \exp(\frac{1}{2}v^2) \sum_{s=0}^{2m} \frac{1}{(n+s)! (n+2m-s)!} \\ &\leq \frac{2m+1}{(n+m)!} \left(\frac{v}{2}\right)^{2(n+m)} \exp(\frac{1}{2}v^2). \end{aligned}$$

LEMMA 5.

$$\int_v^v v^{-(2n+2m+1)} P_{n+1,2m}(v) dv = -\frac{1}{2(2n+2m+1)} v^{-2n-2m} [P_{n,2m}(v) + P_{n+1,2m}(v)].$$

Proof. By definition

$$\begin{aligned} \int_v^v v^{-(2n+2m+1)} P_{n+1,2m}(v) dv &= \int_v^v v^{-(2n+2m+1)} \left[\sum_{s=0}^{2m} (-1)^s J_{n+1+s}(v) J_{n+1+2m-s}(v) \right] dv \\ &= \sum_{s=0}^{2m} (-1)^s \int_v^v v^{-(2n+2m+1)} J_{n+1+s}(v) J_{n+1+2m-s}(v) dv. \end{aligned}$$

The integral under the summation sign can be evaluated by a lemma on cylinder functions [see, for example, Watson (1922, (1), p. 136)]. The above expression then becomes

$$\begin{aligned} \sum_{s=0}^{2m} \frac{(-1)^{s+1}}{2(2n+2m+1)} v^{-(2n+2m)} [J_{n+s}(v) J_{n+2m-s}(v) + J_{n+s+1}(v) J_{n+1+2m-s}(v)] \\ = -\frac{1}{2(2n+2m+1)} v^{-(2n+2m)} [P_{n,2m}(v) + P_{n+1,2m}(v)]. \end{aligned}$$

LEMMA 6.

$$\int_v^v v^{2n+2m+1} P_{n,2m}(v) dv = \frac{1}{2(2n+2m+1)} v^{2n+2m+2} [P_{n,2m}(v) + P_{n+1,2m}(v)].$$

This lemma can be proved by a similar argument as in lemma 5, using Watson (1922, (2), p. 136) in place of Watson (1922, (1), p. 136).

LEMMA 7. $\sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) = J_0(u) \cos u + J_1(u) \sin u.$

Proof. By definition of Q_{2s} and from lemma 3 it follows that

$$Q_{2s}(u) = \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} [J_{2s}(2u \cos \theta) + J_{2s+2}(2u \cos \theta)] \frac{\cos(2s+1)\theta}{\cos \theta} d\theta,$$

and, by applying the addition formula for Bessel functions, we then find that

$$Q_{2s}(u) = \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} \frac{2s+1}{u \cos^2 \theta} J_{2s+1}(2u \cos \theta) \cos (2s+1) \theta d\theta. \quad (\text{A } 5)$$

$$\text{Hence} \quad \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) = \frac{2}{\pi} \sum_{s=0}^{\infty} \int_0^{\frac{1}{2}\pi} \frac{(-1)^s}{u \cos^2 \theta} J_{2s+1}(2u \cos \theta) \cos (2s+1) \theta d\theta. \quad (\text{A } 6)$$

Interchanging the summation and integration sign, which is justified since the series is uniformly convergent in $0 \leq \theta \leq \frac{1}{2}\pi$, and, using the well-known result of Jacobi [see, for example, Watson (1922, (4), p. 22)] that

$$\sum_{s=0}^{\infty} (-1)^s J_{2s+1}(z) \cos (2s+1) \theta = \frac{1}{2} \sin (z \cos \theta),$$

$$(\text{A } 6) \text{ becomes} \quad \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1} Q_{2s}(u) = \frac{1}{\pi} \int_0^{\frac{1}{2}\pi} \frac{\sin (2u \cos^2 \theta)}{u \cos^2 \theta} d\theta. \quad (\text{A } 7)$$

$$\text{Let} \quad F(u) = \frac{1}{\pi} \int_0^{\frac{1}{2}\pi} \frac{\sin (2u \cos^2 \theta)}{u \cos^2 \theta} d\theta. \quad (\text{A } 8)$$

$$\begin{aligned} \text{Then} \quad \frac{d}{du} [uF(u)] &= \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} \cos (2u \cos^2 \theta) d\theta \\ &= \frac{2 \cos u}{\pi} \int_0^{\frac{1}{2}\pi} \cos (u \cos 2\theta) d\theta - \frac{2 \sin u}{\pi} \int_0^{\frac{1}{2}\pi} \sin (u \cos 2\theta) d\theta. \end{aligned}$$

$$\text{Now} \quad \frac{2}{\pi} \int_0^{\frac{1}{2}\pi} \cos (u \cos 2\theta) d\theta = J_0(u),$$

$$\frac{2}{\pi} \int_0^{\frac{1}{2}\pi} \sin (u \cos 2\theta) d\theta = 0,$$

$$\text{so that} \quad \frac{d}{du} [uF(u)] = J_0(u) \cos u.$$

Integrating several times by parts and using the recurrence relations for Bessel functions or otherwise it can easily be verified that the solution of this differential equation subject to the boundary condition $F(0) = 1$ [which is imposed by (A 8)] is

$$F(u) = J_0(u) \cos u + J_1(u) \sin u. \quad (\text{A } 9)$$

Lemma 7 now follows from (A 7), (A 8) and (A 9).

$$\text{LEMMA 8.} \quad \sum_{s=0}^{\infty} (-1)^s V_{n+2s}(v) = W_n(u, v).$$

$$\begin{aligned} \text{Proof.} \quad \sum_{s=0}^{\infty} (-1)^s V_{n+2s}(v) &= \sum_{s=0}^{\infty} (-1)^s \sum_{p=0}^{\infty} (-1)^p \left(\frac{v}{u}\right)^{n+2s+2p} J_{n+2s+2p}(v) \\ &= \sum_{s=0}^{\infty} \sum_{p=0}^{\infty} (-1)^{s+p} \left(\frac{v}{u}\right)^{n+2s+2p} J_{n+2s+2p}(v). \end{aligned}$$

Collecting terms of the same order in J and arranging the series in ascending order lemma 8 follows. The grouping of terms is justified since the series is absolutely convergent.

LEMMA 9.

$$\int_0^{v_0} v V_n(u, v) \exp(-iv^2/2u) dv = u[W_{n+1}(u, v_0) + iW_{n+2}(u, v_0)] \exp(-iv_0^2/2u).$$

Proof.

$$\begin{aligned} \int_0^{v_0} v V_n(u, v) \exp(-iv^2/2u) dv &= \int_0^{v_0} \sum_{s=0}^{\infty} (-1)^s v \left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v) \exp(-iv^2/2u) dv \\ &= \sum_{s=0}^{\infty} \int_0^{v_0} \frac{(-1)^s}{u^{n+2s}} v^{n+2s+1} J_{n+2s}(v) \exp(-iv^2/2u) dv. \quad (\text{A } 10) \end{aligned}$$

The interchanging of the summation and integration sign above is justified, since by (A 4) the series under the integral sign is dominated by the series

$$\sum_{s=0}^{\infty} \frac{v_0 \exp \frac{1}{4} v_0^2}{(n+2s)!} \left(\frac{v_0}{2u}\right)^{n+2s},$$

and is therefore uniformly convergent in the range of integration. The integrals in (A 10) can be evaluated with the help of a result due to Lommel (see, for example, Walker (1904 (10), p. 131) which states, that for $\nu > 0$, $l > 0$,

$$\int_0^r (lx)^\nu J_{\nu-1}(lx) \exp(-ikx^2/2) dx = \frac{l^{2\nu-1}}{k^\nu} [U_\nu(kr^2, lr) + iU_{\nu+1}(kr^2, lr)] \exp(-ikr^2/2)$$

Setting $l = 1$, $x = v$, $r = v_0$, $\nu = n + 2s + 1$, $k = 1/u$ and multiplying by $(-1)^s/u^{n+2s}$, we find

$$\begin{aligned} \frac{(-1)^s}{u^{n+2s}} \int_0^{v_0} v^{n+2s+1} J_{n+2s}(v) \exp(-iv^2/2u) dv \\ = (-1)^s u [V_{n+2s+1}(u, v_0) + iV_{n+2s+2}(u, v_0)] \exp(-iv_0^2/2u). \quad (\text{A } 11) \end{aligned}$$

Lemma 9 now follows on substituting from (A 11) into (A 10) and using Lemma 8.

In conclusion, it is a pleasure to thank Dr E. H. Linfoot for much encouragement and for helpful discussions in the course of this work. I also wish to thank Dr H. H. Hopkins, Mr F. Ursell and Mr P. A. Wayman for useful suggestions. Finally, I wish to acknowledge my indebtedness to the Cambridge University Mathematical Laboratory for assistance with some of the computations and to Miss C. M. Munford for the skill and patience with which she carried them out.

REFERENCES

- Airy, G. B. 1835 *Trans. Camb. Phil. Soc.* **5**, 283.
 Hopkins, H. H. 1949 *Proc. Phys. Soc.* **62 B**, 22.
 Lommel, E. 1868 *Studien über die Besselschen Functionen*. Leipzig: Teubner.
 Lommel, W. 1885 *Abh. Bayer. Akad. Wiss.* **15**, 229. A more accessible, though condensed account of Lommel's treatment is given in: Gray, A., Mathews, G. B. & MacRobert, T. M. 1931 *A treatise on Bessel functions*, 2nd ed. London: Macmillan.
 Nijboer, B. R. A. 1942 Thesis, Groningen.
 Rayleigh, Lord 1881 *Phil. Mag.* **11**, 214. Also *Collected papers*, **1**, 513; **3**, 91.
 Struve, H. 1886 *Mém. Acad. Sci. St-Pétersb.* (7), **34**, no. 5, 1-15.
 Walker, J. 1904 *The analytic theory of light*. Cambridge University Press.
 Watson, G. N. 1922 *A treatise on the theory of Bessel functions*. Cambridge University Press.
 Zernike, F. & Nijboer, B. R. A. 1949 *La théorie des images optiques*. Éditions de la Revue d'Optique, p. 227.

Interferometric studies in diffusion.

II. Influence of concentration and orientation on diffusion in cellulose acetate

BY J. CRANK AND C. ROBINSON

Research Laboratory, Courtaulds Limited, Maidenhead, Berks

(Communicated by A. H. Wilson, F.R.S.—Received 4 July 1950)

[Plates 16 and 17]

The techniques developed in part I are used to study the diffusion of the five penetrants, chloroform, acetone, methylene chloride, water and methyl alcohol into stretched cellulose acetate sheet. The diffusion coefficient—concentration relationship for diffusion at right angles to the direction of stretch is obtained at 25° C for the first three penetrants and also at 40° C for chloroform. In each case a marked increase in diffusion coefficient occurs at a volume concentration of penetrant of about 60 to 70 %. The results for chloroform indicate a mean activation energy of about 6 kcal. Photographs of interferometer fringe systems in the neighbourhood of a corner of the cellulose acetate sheet are shown for each penetrant. Each photograph reveals anisotropic behaviour. This behaviour is due partly to the diffusion coefficient being less in the direction of stretch, than perpendicular to it, and partly to the overall range of concentration being less as a result of the restricted swelling of the polymer in the former direction. The degree of anisotropy is studied quantitatively for chloroform and acetone.

New evidence is provided on the nature of the sharp boundaries seen under the microscope when a penetrant enters a polymer. In particular the middle boundary, which approximately indicates the concentration at which the polymer relaxes and becomes isotropic, is found to occur at a volume concentration of 60 to 70 % for each penetrant examined quantitatively. On inspecting a model of a cellulose acetate molecule it seems likely that at this volume concentration the polymer chains are sufficiently well separated to allow one repeating unit of the polymer chain to rotate about the axis of the chain without hindrance from neighbouring chains. An analogy with second-order transition phenomena is drawn.

INTRODUCTION

In part I (Robinson 1950) an interferometric technique has been described for obtaining the concentration distribution which arises when a penetrant diffuses into a sheet of polymer. This technique was devised in order to study the diffusion coefficient—concentration relationship and other aspects of diffusion behaviour. The development of the technique was illustrated by experiments with the system cellulose acetate—chloroform.

In the present paper convenient formulae are presented by which the concentration distribution can be deduced from an interferometer fringe system, so that the diffusion coefficient—concentration relationship can be calculated by the method due to Matano (1932–3). Further experiments with stretched cellulose acetate and several different penetrants are described and the diffusion coefficient—concentration relationships discussed.

As was mentioned in part I, stretched cellulose acetate sheet was chosen partly in order to study anisotropic swelling and diffusion. Such behaviour is to be expected wherever there is structural anisotropy and is found for example in metals and in

zeolites (Barrer 1941). In the polymer field, anisotropy of equilibrium swelling in artificial and synthetic fibres is a familiar phenomenon, swelling generally being much greater at right angles to the long axis of the fibres and the direction in which the molecules tend to be oriented. Hermans (1949*a, b*) has referred to the importance of such anisotropic behaviour in biological processes. He has also attempted to determine the degree of orientation in 'model' fibres from the anisotropy of swelling.

Hartley (1949) has recently described an extreme case of anisotropic behaviour and some striking phenomena associated with it in a preliminary survey of the diffusion of penetrants into stretched polymer sheet. His observations were made with a polarization microscope and the diffusion process was followed by recording the advance of the sharp inner boundary which accompanies such diffusion. In what follows the data obtained from the interferometric experiments are used in an attempt to obtain a better understanding of the mechanism of this anisotropy. At the same time further light is thrown on the nature of the inner and middle boundaries which are visible under the microscope and which were referred to in part I. Finally, suggestions are put forward concerning the nature of the successive movements of the polymer chains which accompany the progressive dilution.

QUANTITATIVE INTERPRETATION OF A FRINGE SYSTEM

Figure 1*a* shows part of a typical fringe system such as the one shown in figure 5, plate 16. The main qualitative features of the system have already been described in part I. It is proposed here to derive convenient formulae by which the refractive index-distance curve can be deduced from the fringe system in the case where the diffusion occurs between silvered plates inclined to form a wedge-shaped film. Diffusion is supposed to be uni-directional and OX is any line in the direction which intersects two neighbouring fringes in PQ . By considering the difference of optical path between the points P and Q we have

$$\lambda = 2\Delta(\mu l) = 2\bar{\mu}\Delta l + 2\bar{l}\Delta\mu. \quad (1)$$

Here Δl , $\Delta\mu$ are the changes in film thickness and refractive index respectively from P to Q and $\bar{l}$, $\bar{\mu}$ are corresponding mean values over the same range; λ is the wavelength of the light used to produce the fringes. If the wedge angle, denoted by θ , is assumed uniform over the whole area of the fringe system, Δl may be expressed in terms of the fringe spacing f in the pure liquid of refractive index μ_L . Thus in the liquid we have

$$2f\theta\mu_L = \lambda, \quad (2)$$

so that

$$\Delta l = \pm \theta \Delta x \cos \phi = \pm \frac{\lambda \Delta x}{2f\mu_L} \cos \phi, \quad (3)$$

where ϕ is the angle between a linear fringe in the pure liquid and the outer swelling boundary which is perpendicular to the direction of diffusion. Substitution in (1) for Δl from (3) leads to the final relation

$$\Delta\mu = \frac{\lambda}{2\bar{l}} \left(1 \mp \frac{\bar{\mu} \Delta x}{f\mu_L} \cos \phi \right), \quad (4)$$

from which $\Delta\mu$ can be calculated for successive pairs of neighbouring fringes. The quantities Δx , f and ϕ are readily measured from a photograph of the fringe system, while μ_L and λ are usually known or can be measured. The precise treatment of the mean values $\bar{\mu}$ and $\bar{l}$ depends on the accuracy required and on the relative importance of the path differences due to the concentration gradient and the wedge angle. In the present application only a mean value of the thickness l over the whole fringe system was readily obtainable and so this value was adopted for $\bar{l}$ in (4). Actually with the wedge angles used the total change in thickness over the field of diffusion was never

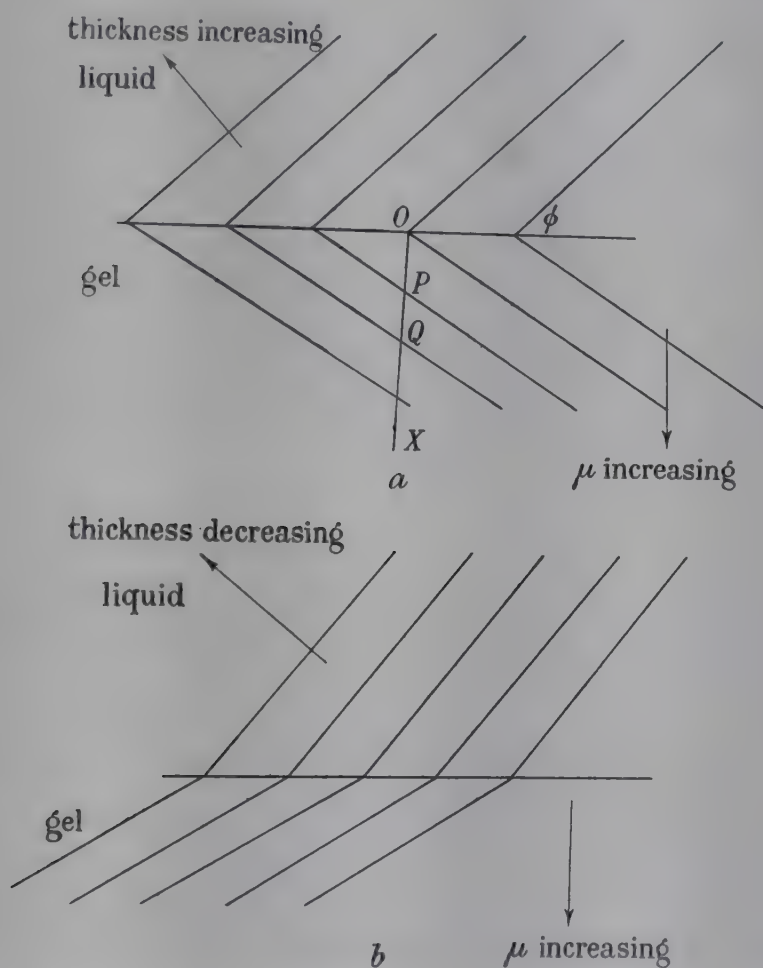


FIGURE 1. Sections of two typical fringe systems.

greater than 2% of absolute thickness. Unless a very high accuracy is required it will be good enough, in calculating $\Delta\mu$ from P to Q , to write $\bar{\mu} = \mu$ at P , and then from Q to R , $\bar{\mu} = \mu$ at Q and so on successively. The process can be started from the first fringe after the so-called meniscus fringes, to which is assigned the value of μ for the fully swollen polymer. An overall check on the calculation and the measurements is provided by a comparison of the final value of μ , obtained by successive addition of all increments $\Delta\mu$ through the whole fringe system, with the measured refractive index for the dry or conditioned polymer as the case may be. A second similar check is provided by the value of μ at the refractive index peak if one exists. Alternatively, this provides a method of obtaining the mean thickness of the film $\bar{l}$. Thus if after correcting for the wedge angle there are N fringes between two points,

e.g. the fully swollen polymer and the peak, for which there is a known refractive index difference of $\Delta\mu_N$, then $\bar{l}$ is given by

$$\bar{l} = N\lambda/2\Delta\mu_N. \quad (5)$$

The positive and negative signs in (4) arise because the wedge thickness may increase or decrease along the direction in which μ increases. The correct sign to use in any given system can be determined from the form of the fringes, simply by remembering that the fringes are contours of constant optical path difference and therefore that along a fringe an increase in μ must be accompanied by a decrease in film thickness. Thus, if in figure 1*a*, μ increases along OX , the fringe pattern has the general form shown in the region of the outer swelling boundary if the wedge thickness increases in the direction shown in the figure. If the wedge thickness decreases in this direction, however, the fringes will be as shown in figure 1*b*. For a fringe system such as in figure 1*b* a positive sign is used in equation (4), and for figure 1*a* a negative sign.

Special cases of the general formula (4)

(i) *Zero wedge angle or wedge angle fringes parallel to direction of diffusion*

In this case either $f = \infty$ or $\cos \phi = 0$ so that (4) reduces to

$$\Delta\mu = \lambda/2\bar{l}, \quad (6)$$

and $\bar{l}$ is now the constant thickness of the film.

In this simple case N in (5) becomes the number of observed fringes, there being no wedge angle correction, and combining (5) and (6) we have simply

$$\Delta\mu = \Delta\mu_N/N. \quad (7)$$

(ii) *Small wedge angle or almost uniformly spaced fringes*

In actual fact the wedge correction varies as the distance in the direction of diffusion. If the wedge angle is not large, however, or if the spacing of the fringes is not too non-uniform, then the wedge angle correction can be assumed to be proportional to fringe number rather than to distance. From (1) we have

$$\Delta\mu = \frac{\lambda}{2\bar{l}} - \frac{\bar{\mu}\Delta l}{\bar{l}}, \quad (8)$$

$$\text{and from (5)} \quad \frac{\lambda}{2\bar{l}} = \frac{\Delta\mu_N}{N} = \frac{\Delta\mu_N}{n - n_w}, \quad (9)$$

where n is the number of observed fringes and n_w the number of wedge-angle fringes in the range over which μ changes by $\Delta\mu_N$. Also with our present assumption that the wedge-angle correction is equally distributed between the fringes

$$\Delta l = n_w\lambda/2n\mu_L, \quad (10)$$

so that, substituting from (10) and (9) in (8),

$$\Delta\mu = \frac{\lambda}{2\bar{l}} \left(1 - \frac{n_w\bar{\mu}}{n\mu_L} \right) = \frac{\Delta\mu_N}{n - n_w} \left(1 - \frac{n_w\bar{\mu}}{n\mu_L} \right). \quad (11)$$

The ratio $\bar{\mu}/\mu_L$ arises because the number of wedge-angle fringes n_w is estimated from the pure liquid region of refractive index μ_L . For present purposes the ratio can be taken as unity and we have the final simple formula

$$\Delta\mu = \Delta\mu_N/n. \quad (12)$$

In a system where this approximation is good enough the change of refractive index per fringe is simply the total change of refractive index divided by the total number of observed fringes between two points of known refractive index. Formula (12) was used in calculating most of the results presented in this paper.

CALCULATION OF DIFFUSION COEFFICIENT

The refractive index-concentration data of figure 12 of part I and the refractive index—distance data calculated in the manner just described, are combined to obtain a concentration—distance curve, examples of which are shown in figure 6 of part I. If it is assumed that at a given temperature the diffusion coefficient does not depend on any factor other than concentration, the nature of this concentration dependence can be deduced from a single concentration—distance curve observed at any one time. The method appears to have been applied first by Matano (1932–3) to the inter-diffusion of two metals. With the above assumption and for certain boundary conditions which include those of the present experiments, Boltzmann (1894) showed that the concentration may be expressed as a function of a single variable y where

$$y = x/t^{\frac{1}{2}}. \quad (13)$$

Here x is the space co-ordinate in the direction of diffusion and t is time.

The diffusion of the penetrant under the experimental conditions described is governed by the normal diffusion equation in one dimension

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x} \left(D \frac{\partial C}{\partial x} \right), \quad (14)$$

together with the initial conditions

$$C = C_0, \quad x < 0, \quad t = 0, \quad (15)$$

$$C = 0, \quad x > 0, \quad t = 0, \quad (16)$$

where C is the concentration of penetrant. On substituting the new variable y , equation (14) reduces to the ordinary differential equation

$$-\frac{y}{2} \frac{dC}{dy} = \frac{d}{dy} \left(D \frac{dC}{dy} \right), \quad (17)$$

which on integration with respect to y becomes

$$-\frac{1}{2} \int_0^{C_1} y dC = \left(D \frac{dC}{dy} \right)_{C=0}^{C=C_1} = \left(D \frac{dC}{dy} \right)_{C=C_1}, \quad (18)$$

since $DdC/dy = 0$ when $C = 0$. Here C_1 is any value of C between 0 and C_0 . Finally, by rearrangement of (18) and reintroducing x and t we have

$$D_{C-C_1} = -\frac{1}{2t} \frac{dx}{dC} \int_0^{C_1} x dC. \quad (19)$$

Since $(DdC/dy)_{C=C_0} = 0$ also, it follows from (18) that

$$\int_0^{C_0} ydC = \int_0^{C_0} xdC = 0. \quad (20)$$

In order that the boundary conditions shall be satisfied the origin from which x is measured must be such that (20) is satisfied. In a constant volume system (20) becomes a conservation of mass relation and it is clear that it is satisfied if x is measured from the initial position of the boundary between penetrant and polymer at time $t = 0$. In all systems with which we are concerned in this work it is a good enough approximation to neglect any small volume change which may occur on mixing.

The procedure therefore is to plot the concentration—distance curve for a known time on squared paper, to locate the line $x = 0$ by use of (20), and then to evaluate D at various concentrations C_1 from (19). The integrals were obtained by counting squares and the gradients by drawing tangents.

The curves of figure 6, part I support the assumption that C is a function of $x/t^{1/2}$ only, i.e. that D is a function of C only. It is a corollary of this that all concentration—distance curves at successive times should intersect in one point, and that this point should lie on the line defining the initial boundary between penetrant and polymer, since, for this line, $x/t^{1/2}$ is constant and zero and therefore the concentration is constant for all times at $x = 0$. The experimental evidence shows that this is so within the expected limits of error.

Intrinsic diffusion coefficient

The diffusion coefficient calculated in the manner just described is a mutual coefficient for the system, e.g. chloroform—cellulose acetate. Measured in this way the diffusion coefficient for chloroform in cellulose acetate is identically equal to that for cellulose acetate in chloroform. It has been shown (Hartley & Crank 1949; Darken 1948) that this equality results from the condition of constant overall volume of the system on mixing, and that it can be reconciled with the expected large difference in diffusion rates of a small chloroform molecule and a large polymer molecule in the same environment by assuming a compensating mass-flow of the whole system to take place. In a solvent—polymer system the movement of the polymer is due almost entirely to this mass-flow.

The rate of transfer of chloroform measured by the mutual diffusion coefficient is therefore the result of two effects, that is diffusion of the chloroform with respect to the mass flow, and the transfer resulting from the mass flow itself. From some points of view it is more convenient to have a diffusion coefficient which measures the rate of transfer of chloroform relative to the mass flow i.e. relative to its environment. Hartley & Crank (1949) have referred to this as an intrinsic diffusion coefficient and have shown that if the intrinsic diffusion coefficient of the polymer is negligible compared with that of the penetrant the intrinsic coefficient of the penetrant e.g. chloroform, denoted by $\mathcal{D}$, is related to the mutual coefficient D by the expression

$$\mathcal{D} = D/(\text{volume fraction of polymer}). \quad (21)$$

We shall see later when values of the diffusion coefficients are given that it is not unreasonable to use the approximate relation (21). It seems likely that such a quantity as the activation energy of diffusion, for example, will be more revealing as regards the mechanism of the diffusion if calculated on the basis of $\mathcal{D}$ rather than of the mutual coefficient D .

DIFFUSION COEFFICIENTS FOR CELLULOSE ACETATE AND CHLOROFORM PERPENDICULAR TO THE DIRECTION OF STRETCH

We may now consider the diffusion coefficients which have been calculated by this method from the experimental data reported in part I.

Concentration dependence

Figures 2 and 3 show the dependence of the mutual diffusion coefficient D on concentration of chloroform at temperatures of 25 and 40° C respectively.

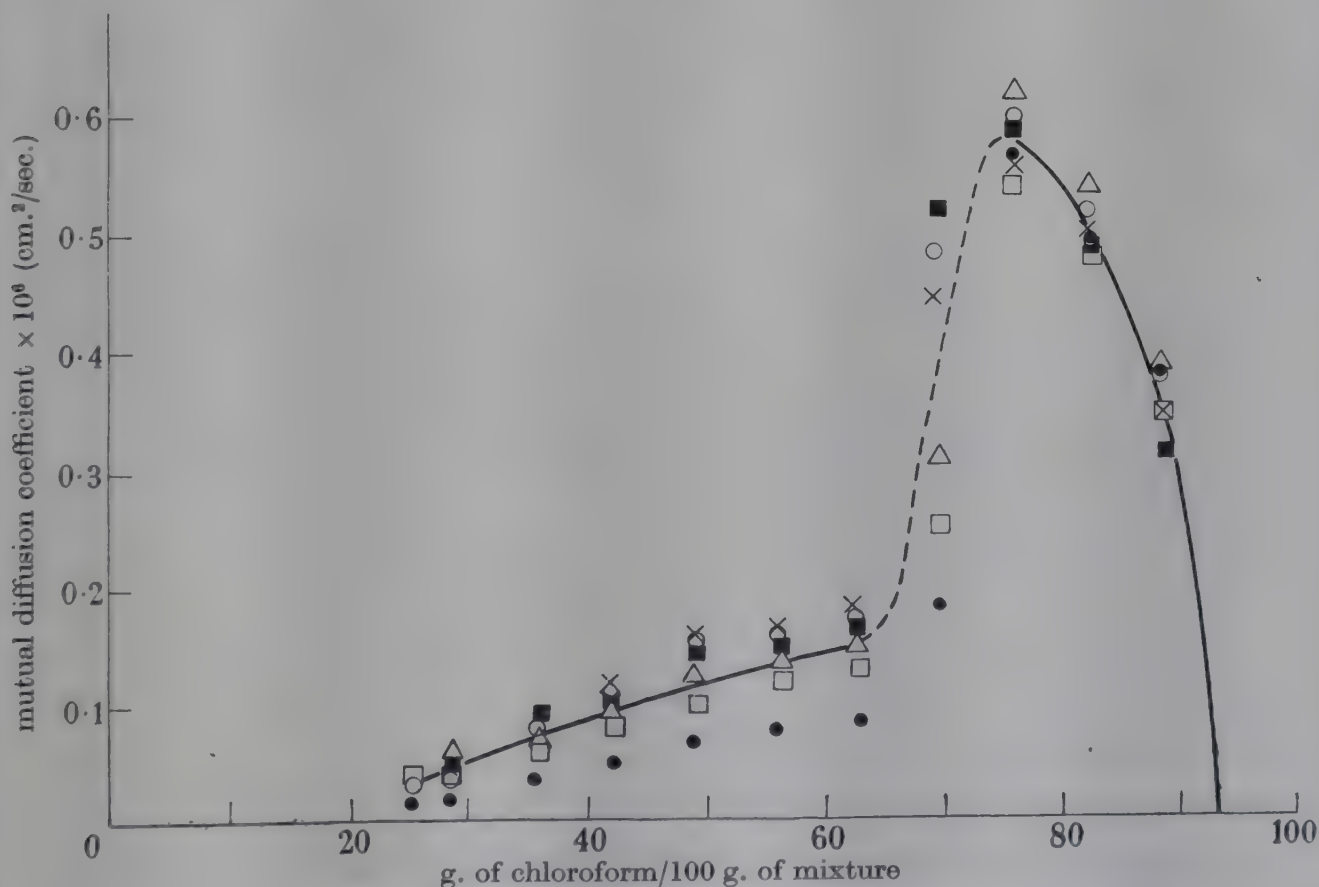


FIGURE 2. Mutual diffusion coefficient for chloroform and cellulose acetate (easy way) at 25° C.

	experiment no.	time in min.
●	38	200
○	41	120
△	41	180
□	41	256
■	44	154
×	44	92
—	mean curve	

In each case results are shown from three different experiments on nominally identical specimens of cellulose acetate and a number of different times. The full line results from taking the arithmetic mean of the different values at each concentration so that the behaviour of the mean can be examined and also some idea of its significance is afforded by the scatter of the individual points. The portions of the curves shown broken are referred to below. The measurements on which these values

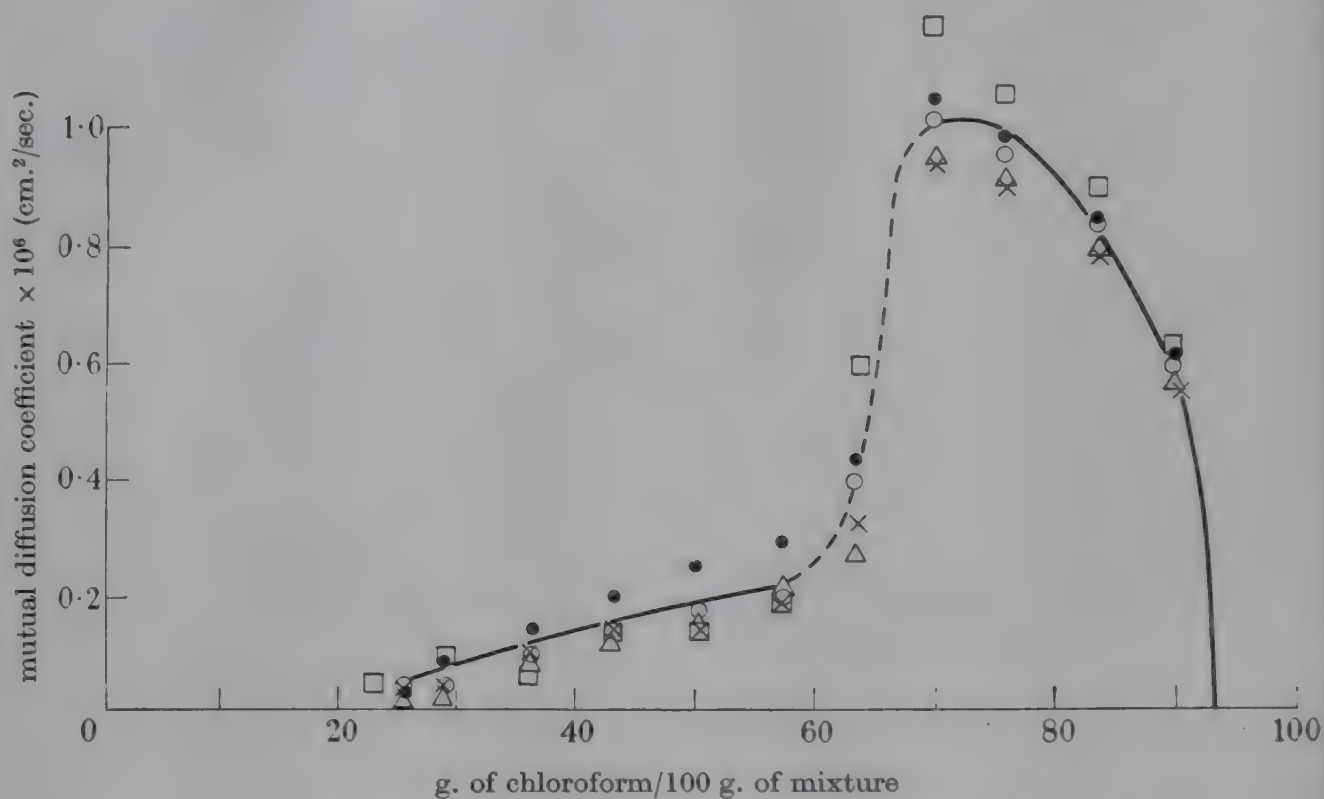


FIGURE 3. Mutual diffusion coefficient for chloroform and cellulose acetate (easy way) at 40° C.

	experiment no.	time in min.
●	39	81
×	39	145
○	40	90
△	40	145
□	67	148
—	mean curve	

of diffusion coefficient are based relate to the penetration of chloroform into cellulose acetate perpendicular to the direction of orientation of the cellulose acetate molecules. Because the cellulose acetate was initially conditioned to a uniform concentration of chloroform of 22 % no measurements of D are available below this concentration of chloroform. (As in part I concentration is expressed as percentage by weight). This is therefore the lower limit of concentration on figures 2 and 3. The infinite concentration gradient over the concentration range between fully swollen acetate and pure chloroform, i.e. between chloroform concentrations of 93 % and 100 % (see figure 6, part I) must be associated with a zero diffusion coefficient, since in equation (19) the gradient dx/dC is zero while the integral is finite. Thus the mean curves of figures 2 and 3 each drop to zero at a chloroform concentration of 93 %.

The main features of the curves of figures 2 and 3 are: (a) over the low concentration range D increases with chloroform concentration increasing, (b) over the high concentration range D decreases as chloroform concentration increases, (c) in the region of chloroform concentration of 60 to 70 % the value of D increases very rapidly by a factor of four or five. This sudden increase in diffusion coefficient occurs in the region of what we have called the middle boundary. Clearly the errors of

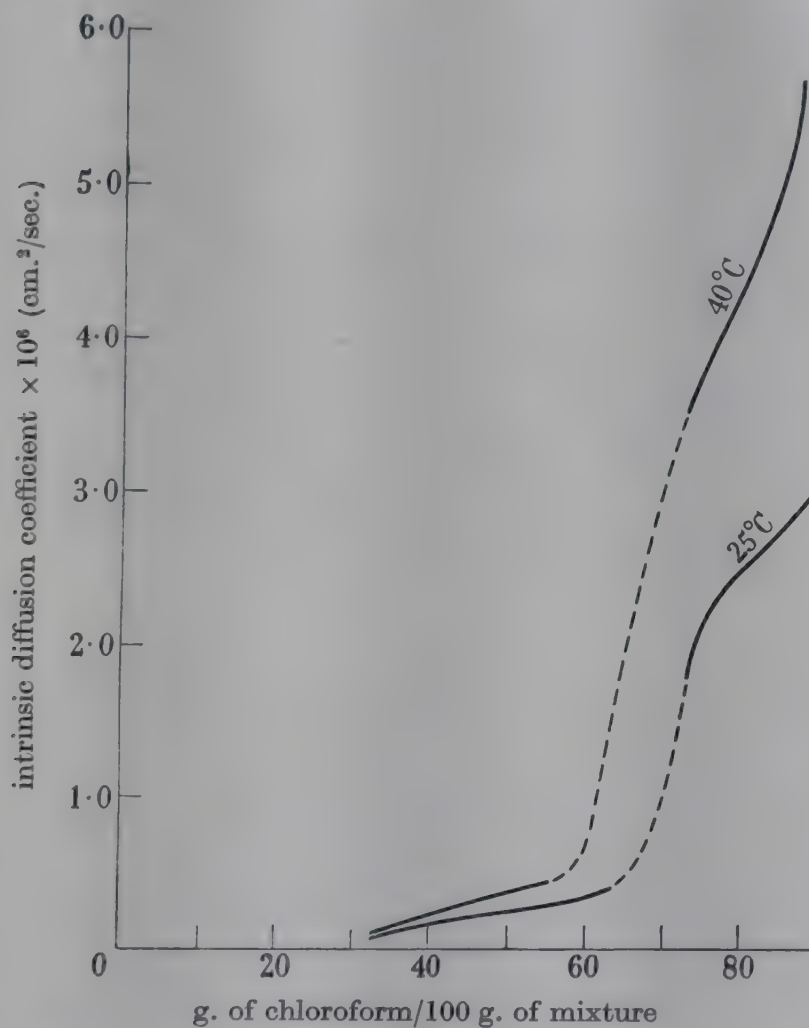


FIGURE 4. Intrinsic diffusion coefficients for chloroform and cellulose acetate at 25 and 40° C.

measurement and calculation in this region are relatively large and so the detailed shape of the D - C curve is less certain in this region than elsewhere. For this reason a broken line is used in figures 2 and 3. If there is a true discontinuity in the gradient of the concentration-distance curves of, say, figure 6, part I at the middle boundary this will be reflected in a discontinuity in the value of D at the same concentration, but the present data do not allow us to distinguish with certainty between a discontinuous increase and a rapid but continuous one.

In figure 4 the concentration dependence of the intrinsic diffusion coefficients calculated from the mean mutual coefficients at the two temperatures by use of equation (21) is shown. At each temperature the concentration dependence of the intrinsic coefficient is similar to that of the mutual coefficient except at high chloroform concentrations where the values of $\mathcal{D}$ continue to increase as chloroform

concentration increases. Consideration of the relative sizes of the chloroform and cellulose acetate molecules suggests that the assumption on which (21), and therefore the curves of figure 4, are based is a reasonable one. Some other quantity, e.g. the velocity of the mass flow (Hartley 1946) must be measured as well as the mutual diffusion coefficient if more reliable intrinsic diffusion coefficients are to be obtained.

Temperature dependence

Table 1 shows the ratio of the intrinsic diffusion coefficient at 40° C to that at 25° C for several concentrations of chloroform. The values of the activation energy, E , for diffusion shown in table 1 suggest that this is somewhat greater at high than at low concentrations of chloroform. The errors in E , particularly at low concentrations, are however considerable and so too much reliance should not be placed on the concentration dependence. The mean order of magnitude is interesting being of the order of a single hydrogen bond.

TABLE 1

percentage by weight of chloroform	$\mathcal{D}_{40^{\circ}\text{C}}/\mathcal{D}_{25^{\circ}\text{C}}$	mean activation energy E (kcal./mol.)
35	1.56	5.6
40	1.55	5.5
45	1.55	5.5
50	1.56	5.6
55	1.62	6.0
75	1.70	6.6
80	1.72	6.8
85	1.77	7.1

ANISOTROPIC BEHAVIOUR

So far we have only considered the diffusion of penetrant across the direction of orientation of the polymer molecules. Hartley (1949) has recently suggested that the diffusion behaviour of stretched cellulose acetate sheet might be expected to be anisotropic and has given experimental evidence of such behaviour for a number of penetrants. He observed the rate of advance of the sharp inner boundary visible under a polarization microscope. By using a circular piece of stretched cellulose acetate sheet he was able to compare the rates of advance of this boundary along the two diameters, parallel and perpendicular to the direction of stretch and thus obtain a measure of anisotropy.

The fringe systems obtained in our experiments, by providing much more complete information about the behaviour of the penetrants, allow a considerable extension and modification of the conclusions which Hartley was able to draw from his exploratory experiments. A convenient way to study anisotropic behaviour with the interferometer is to photograph the fringe system in the neighbourhood of a corner of the stretched cellulose acetate sheet, so that the behaviour in the easy or difficult ways can be observed simultaneously. Such corner photographs are shown in figures 5 to 9, plates 16 and 17, the penetrants being chloroform, acetone, methylene chloride, water and methyl alcohol, respectively. The direction of stretch in all the photographs is at right angles to the bottom of the page. When discussing the

anisotropy, we shall, for convenience, follow Hartley's (1949) nomenclature and call the direction perpendicular to the direction of stretch the easy way, and that parallel to it, the difficult way. Acetone is the only one of these penetrants which is a solvent for cellulose acetate and is usually stated to be miscible with it in all proportions. Marked anisotropic behaviour is seen in all five cases, the disposition of the fringes about the corner being far from symmetrical. The distance between the outermost and innermost limits of the region in which the fringes are noticeably distorted because of the diffusion, is much greater in the easy than in the difficult way. Further, the number of fringes (after allowing for the wedge angle fringes) between the undisturbed polymer and the pure penetrant is less for the difficult than for the easy way. Thus taking the photograph for conditioned chloroform as an example (figure 5, plate 16), some thirty fringes are visible in the easy direction compared with only five in the difficult direction (a similar photograph for chloroform and unconditioned cellulose acetate also showed some penetration in the difficult direction). These five fringes are due to a penetration of the polymer by chloroform, but so far as can be judged this penetration has occurred with little or no outward swelling of the polymer. This restricted swelling is in fact perhaps the most noticeable feature of the anisotropy. The five fringes represent a change in refractive index of 0.0048 which means that the concentration of chloroform at the outer edge of the sheet in the difficult direction is only 34 % as compared with 93 % in the easy direction. This is roughly the concentration at which the refractive index maximum occurs and the behaviour is consistent with the idea that penetration with little or no swelling can occur below the chloroform concentration at which the free space postulated by Hermans is fully occupied.

It is at once apparent that the anisotropic behaviour is not entirely due to the diffusion coefficient being greater in one direction than the other, but is due, at least in part, to the fact that the surface concentration is lower in the difficult than the easy way. In considering diffusion of a penetrant into a solid it is usually assumed that a surface molecular layer of the solid instantaneously takes up its equilibrium concentration of penetrant and that subsequent diffusion into the solid takes place from this layer. The rate of penetration of any given concentration of penetrant depends on the diffusion coefficient and also on the surface concentration. This is a concept familiar in steady-state diffusion through membranes where the permeability is written as the product of diffusion coefficient and solubility. Thus even if in the systems in which we are interested, the diffusion coefficient was independent of concentration, and had the same value in both easy and difficult directions, penetration the difficult way would be relatively slow because of the reduced surface concentration. We already know, however, that, in the easy direction, the diffusion coefficient decreases considerably at low concentrations of chloroform. If the concentration dependence in the difficult direction is comparable, the rate of penetration will be even more reduced in the difficult direction because only the low values of the diffusion coefficient at the lower concentrations will be operative.

In view of these considerations it is clearly of interest to attempt to calculate the diffusion coefficient in the difficult direction. This is possible in the case of chloroform where sufficient quantitative data are available. In principle the calculation of the

diffusion coefficient in the difficult direction is the same as in the easy direction. In practice, however, since far less fringes are visible in the difficult direction the accuracy with which the concentration—distance curve and hence the diffusion coefficient—concentration relation can be obtained is much less than in the easy direction. For this reason the results do not justify the presentation of a complete diffusion—concentration curve for the difficult way. The coefficient has a mean value of about 10^{-8} cm.²/sec. and there is some indication that its value increases with concentration increasing. The figure of 10^{-8} cm.²/sec. is a mean value over the concentration range from 22 to 34 % chloroform. The mean value over the corresponding range in the easy direction is about 5×10^{-8} cm.²/sec. Thus for the same concentration the diffusion coefficient is roughly 5 times less for the difficult than for the easy direction.

The interferometer photographs therefore reveal some anisotropy of diffusion coefficient, and this is sufficient to account for a factor of $\sqrt{5}$ in the anisotropy of penetration. Taking the penetration of a certain concentration to mean the distance this concentration has moved inwards from the initial position of the boundary between chloroform and polymer, the observed anisotropy of penetration for a concentration of 25 % chloroform, say, is roughly 4 times. Thus $2\frac{1}{2}$ times are due to the difference in diffusion coefficient and the remainder due to the reduced swelling in the difficult direction.

In the paper by Hartley (1949) two possible explanations of anisotropy of penetration were put forward. There was first the possibility that in an oriented polymer film the movements of segments of the polymer chains by thermal vibration are less likely to facilitate diffusion of small molecules parallel to the length of the chain than across it. This leads to anisotropy of what we may call the purely random diffusion movement of penetrant molecules. The second suggested mechanism is based on the well-known fact (Mark 1943) that it is easier to stretch mechanically an oriented film perpendicular to the direction of orientation than along it. Now in general the entrance of penetrant will lead to stretching of the polymer as it swells to accommodate the additional volume. This stretching or mass flow as it has been called will occur more readily perpendicular to, rather than along, the direction of orientation and this may be expected to result in a slower rate of penetration in the latter direction.

Our results for chloroform show that both these mechanisms play some part in the anisotropy. Thus the diffusion coefficient is reduced in the concentration range for which penetration is observed. There is also an almost total restriction of mass flow (swelling) in the difficult direction, and this contributes to the anisotropy of penetration at concentration less than 34 % chloroform. At higher concentrations of chloroform the restriction of mass flow appears to be the main cause of anisotropy. The inability of these specimens of cellulose acetate to stretch further in the direction of orientation is not surprising in view of the fact that they were originally oriented by stretching in hot methyl alcohol to the greatest extent possible without breaking. In our experiments the interpretation is relatively straightforward because the mass flow in the difficult direction is negligible. If the specimens were less highly oriented initially, however, so that some mass flow were possible even in the direction of orientation, then the anisotropy of the diffusional movements of the penetrant

molecules and that of the mass flow would be combined and would interact with each other in a rather complex way to produce the overall anisotropic behaviour. Further investigation of this more general case is outside the scope of the present work.

The anisotropy of penetration clearly depends very much on the particular concentration observed, the extreme case being reached for concentrations of chloroform greater than 34 % for which, as we have seen, there is no detectable penetration or swelling in the difficult direction. In this connexion it should be realized that the interferometer can show penetration, by disturbance of the fringe system, in cases where no penetration would be visible under the microscope even with the help of polarized light. Thus in figure 5, plate 16, the fringes show some penetration of chloroform the difficult way, but no inner advancing boundary is visible under the

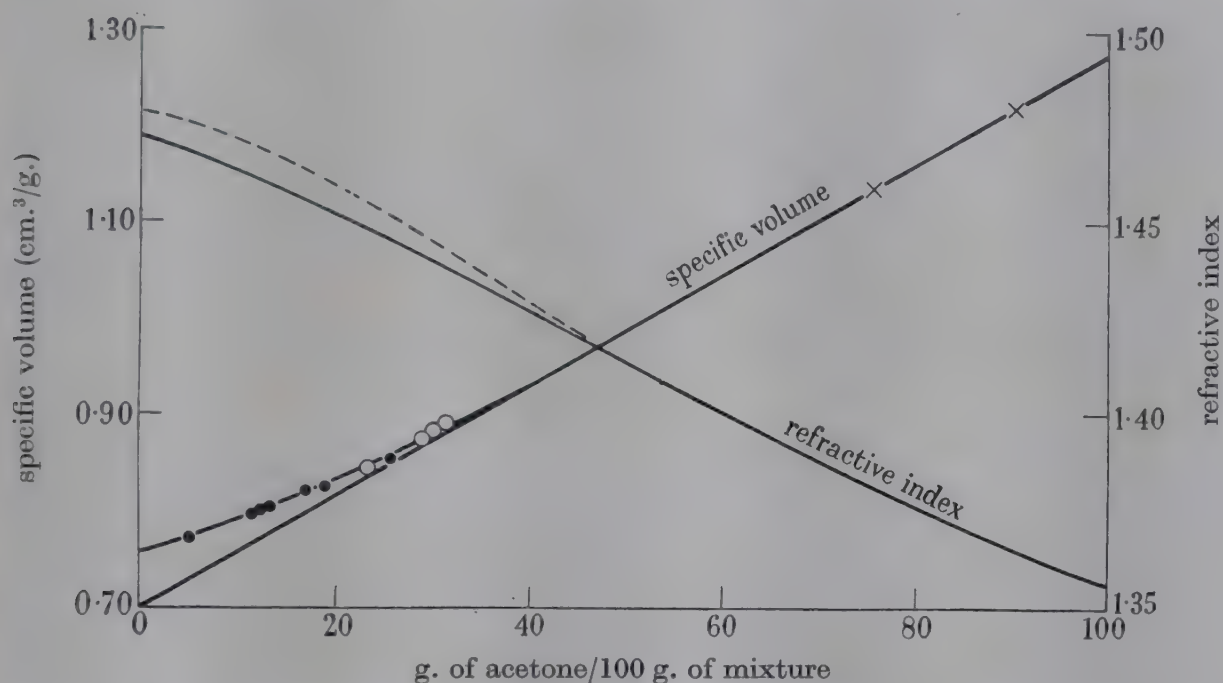


FIGURE 10. Specific volume and refractive index curves for acetone—cellulose acetate at 25° C; — refractive index for electric vector perpendicular to direction of stretch; -- refractive index for electric vector parallel to direction of stretch; ●, density column; ○, immersion method; ×, density bottle.

microscope. Thus from the microscope observation alone we should conclude that we had an extreme case of anisotropy with no penetration in the difficult direction. This explains why Hartley concluded from his observations with the microscope that with all normal solvents and swellers no penetration occurs in the difficult direction.

The diffusion coefficients for the two directions were also calculated in the case of acetone. A specific volume curve was determined as for chloroform and a refractive index curve was calculated from this (figure 10). Since the refractive index of acetone is greater than that of cellulose acetate there is no maximum in this calculated curve. This is consistent with the observation that no refractive index peak is seen in the fringe system. The calculated refractive index curve was used for calculating refractive index gradients. (At present experiments with cellulose acetate conditioned to contain various concentrations of acetone have not been carried out, and

so it has not been possible to construct a refractive index curve directly from the interferometer experiments as in the case of chloroform.)

For the easy direction considerable difficulty was found in resolving all the fringes on the polymer side of the middle boundary as the fringe spacing there is very small. It was necessary to use two negatives to cover the whole system and to reduce the exposure as much as possible to prevent blurring of the fringes. Very accurate focusing was necessary which was obtained by observing a piece of plain glass in the focal plane of the camera with a low-power microscope. Eventually the fringe system shown in figure 6, plate 17 was completely resolved. Part of this system is plotted in figure 11. Seventy-seven fringes were counted which agreed with the expected

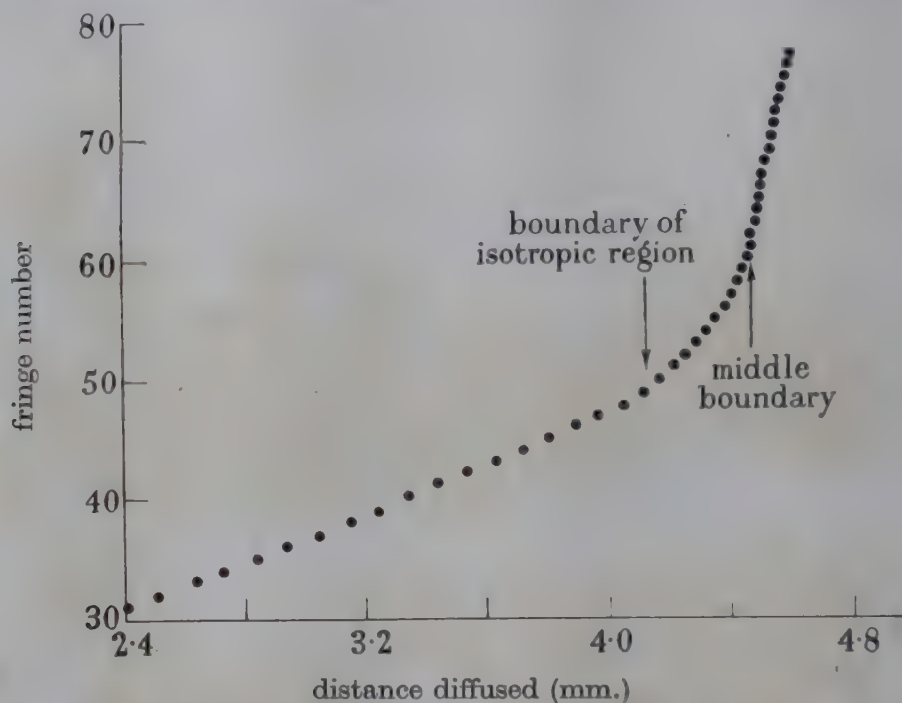


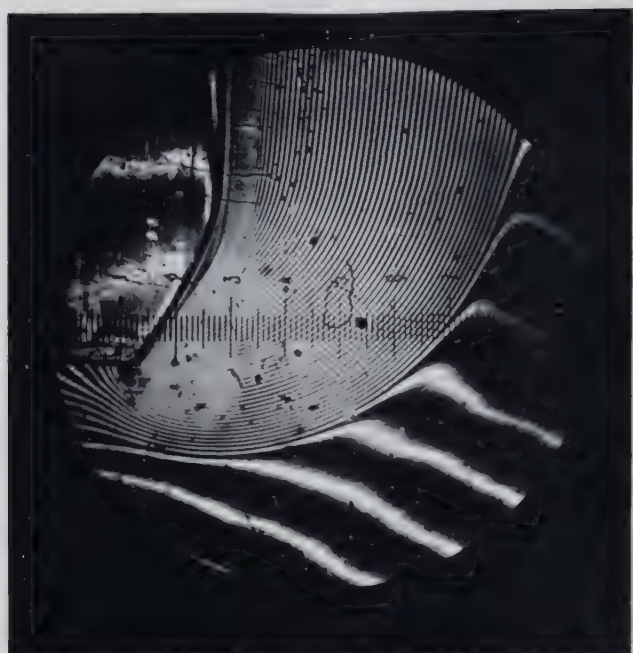
FIGURE 11. Part of the fringe number—distance plot for acetone (easy way) at 25° C and 1435 minutes.

number calculated from the thickness of the specimen and the refractive index difference between polymer and penetrant, after the usual corrections had been made. The fringe system can therefore be used to determine the complete diffusion coefficient—concentration curve from undiluted polymer to pure solvent. It is interesting to notice how well defined the outer boundary has remained when using a solvent penetrant, in spite of diffusion being observed while the plane of the cellulose acetate was vertical. With unconditioned cellulose acetate there is no penetration the difficult way, but some eleven fringes are to be seen spreading into the solvent region (figure 6, plate 17). The fringe system shows the complete miscibility in the easy direction, but the smaller number of fringes in the difficult direction shows that the highest concentration of cellulose acetate in solution in acetone is only 15 %.

In figure 12 the calculated diffusion coefficients for acetone are given. The diffusion coefficients for the difficult way are shown for the limited concentration range covered in this direction. It will be seen that there is no significant difference between the



FIGURE 5. Experiment 44. Cellulose acetate conditioned to 20.3 % chloroform; chloroform entering from right; 25° C; thickness 0.195 mm.; time 251 min.; magnification $\times 23.2$.



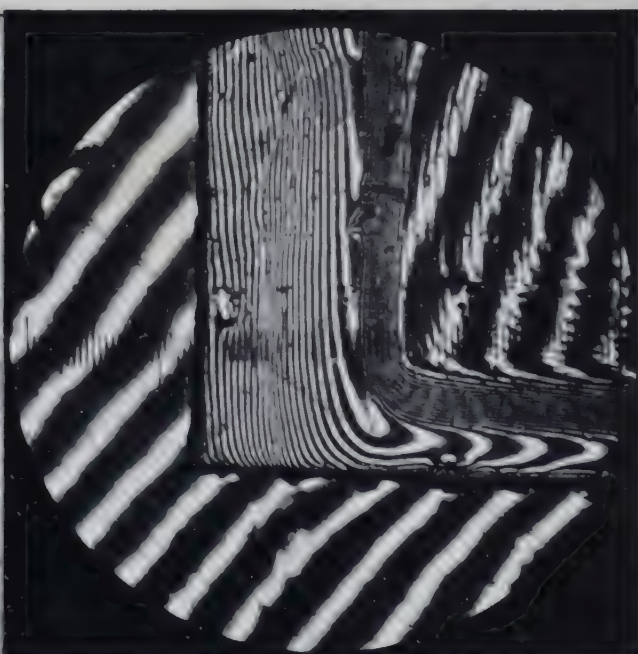
6



7



8



9

FIGURE 6. Experiment 57. Unconditioned cellulose acetate; acetone entering from right; 25°C ; thickness 0.165 mm. ; time 38 min. ; magnification $\times 22.5$.

FIGURE 7. Experiment 103. Unconditioned cellulose acetate; methylene chloride entering from right; 25°C ; thickness 0.180 mm. ; time 125 min. ; magnification $\times 12.7$.

FIGURE 8. Experiment 124. Unconditioned cellulose acetate; water entering from left; 25°C ; thickness 0.180 mm. ; time 170 min. ; magnification $\times 25.3$.

FIGURE 9. Experiment 125. Unconditioned cellulose acetate; methyl alcohol entering from left; 25°C ; thickness 0.200 mm. ; time 1100 min. ; magnification $\times 24.6$.

diffusion coefficients in the two directions. The anisotropic behaviour with acetone, unlike that with chloroform, is therefore, at these concentrations, entirely due to a difference in the mean overall concentration gradients for the two directions, the polymer being completely miscible in one direction and having only a limited

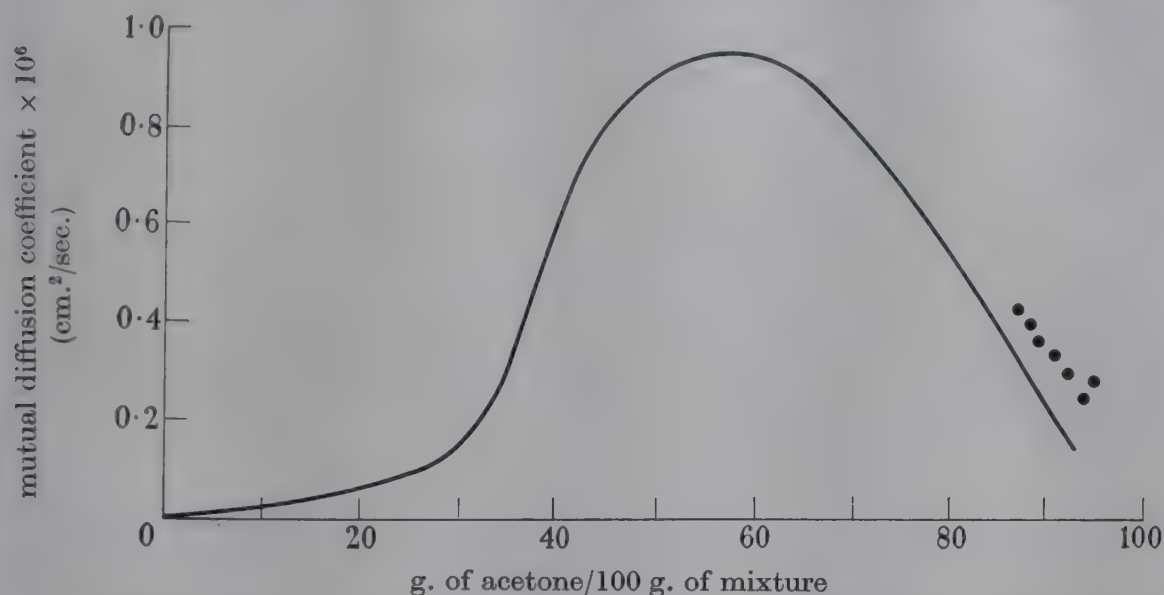


FIGURE 12. Mutual diffusion coefficients for acetone and cellulose acetate at 25° C; —, easy way; ..., difficult way. (In this experiment a small amount of water was present in the acetone.)

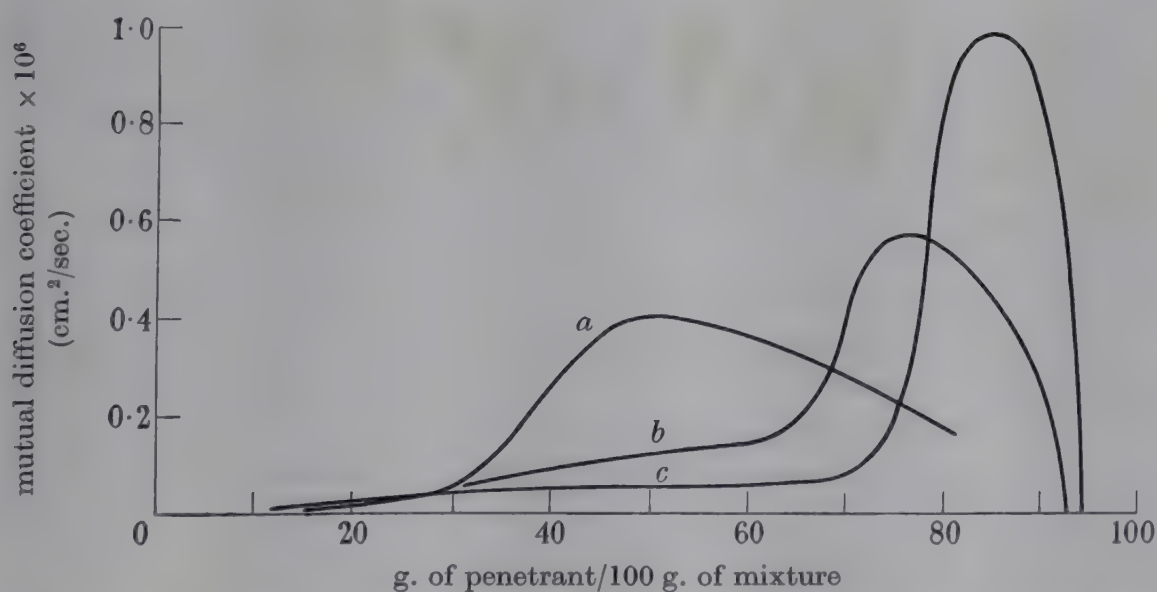


FIGURE 13. Mutual diffusion coefficients for three penetrants and cellulose acetate (easy way) at 25° C. The curve shown for chloroform is the mean curve from figure 2. *a*, acetone; *b*, chloroform; *c*, methylene chloride.

'solubility' in the other. On the other hand, an alternate point of view is that for concentrations higher than those for which results are given, the diffusion coefficient is zero or very low. In the latter case more fringes would eventually become visible.

The experiment described above is the only one obtained so far with acetone in which every fringe is resolved and in which at the same time a satisfactory observation of the difficult way diffusion can be made sufficiently far from the corner to avoid 'corner

effects'. Unfortunately, no precautions were taken to prevent the acetone from taking up water from the atmosphere and it is probable that a small amount of water was present. (The presence of the water would not greatly affect the refractive index.) The diffusion coefficient curve for the easy way obtained from an experiment with dry acetone is included in figure 13 together with the corresponding curves for chloroform and methylene chloride. Comparing figures 12 and 13 it would seem that the presence of the water has considerably increased the diffusion coefficient. The 'solubility' in the difficult direction with dry acetone (although not accurately determined) was found to be even less and the anisotropic behaviour was consequently even more marked than when the acetone contains some water.

A similar experiment using cellulose acetate conditioned to contain 16 % acetone, showed some penetration as well as solubility in the difficult direction.

With methylene chloride (figure 7, plate 17) the anisotropic behaviour is even more marked, little or no penetration taking place in the difficult direction with dry, unconditioned cellulose acetate.

THE VISIBILITY OF BOUNDARIES UNDER THE MICROSCOPE

We have said earlier that in some systems three boundaries are visible under the microscope and these have been referred to as the inner, outer and middle boundaries. It is proposed in this section to summarize the new evidence provided by the interferometer fringe systems on the nature and cause of these boundaries.

As soon as the interferometer was used for following uniplanar diffusion, it became clear that the appearance of a boundary line in the microscope did not necessarily indicate a discontinuity in the refractive index—distance curve as was at first thought. Thus when, for example, chloroform is diffusing into conditioned cellulose acetate there is no suggestion of a discontinuity in the refractive index—distance curve. Every fringe can be resolved in the neighbourhood of the inner boundary, although this is clearly visible under the microscope. The concentration—distance curves of figure 6, part I show that the boundary is associated with a rapid change in concentration gradient, rather than with a discontinuity in concentration. This conclusion has also been reached on theoretical grounds (Crank & Henry 1949). Thus concentration—distance curves, calculated for a type of diffusion coefficient known to give rise to a sharp inner boundary, showed nothing approaching a discontinuity in concentration but the same tendency to an abrupt termination of the concentration—distance curves as we have just noted. One such calculation was carried out for the chloroform—polystyrene system. The diffusion coefficient used in the calculation was measured by an alternative method (Crank & Park 1949) in which only the overall rate of absorption of penetrant is observed and not its distribution within the polymer.

In some cases the inner boundary observed under the microscope does not mark the furthest advance of penetrant, since it is sometimes preceded by fringes which curve gradually and continuously into the wedge angle fringes. Also, as we have previously noted, when diffusion of chloroform parallel to the direction of stretch of the cellulose acetate is observed no boundary is seen under the microscope though several fringes are visible in the interferometer merging gradually into the wedge

angle fringes. The precise position of the inner boundary if one is seen, will depend on the shape of the concentration—distance curve in that region and this, of course, is a reflexion of the concentration dependence of the diffusion coefficient there. Thus no boundary is seen if the diffusion coefficient is constant but only if the coefficient increases in certain ways with concentration increasing. The position of the boundary will also depend on the behaviour of the refractive index—concentration relation.

Similar conditions give rise to the outer boundary seen, for example, with acetone as penetrant. When the penetrant is a sweller only and not a solvent, an outer boundary is again visible marking the limit of miscibility. In this case, however, it may be due to the discontinuity in the concentration (see the concentration—distance curves of figure 6, part I).

What we have referred to as the middle boundary is a further example of a boundary caused by a sufficiently rapid change of refractive index gradient, rather than a discontinuity or rapid change in concentration. Thus with cellulose acetate conditioned to 22 % chloroform, for example, a middle boundary can be seen under the microscope and also every fringe is visible in this region in the interferometer.

From these observations it seems reasonable to postulate that a boundary of this kind occurs where $d^2\mu/dx^2$ is a maximum, and it appears that this maximum must exceed some lower limiting value. Thus if the rate of change of refractive index gradient in the region of the middle boundary is decreased (for instance in the later stages of a diffusion experiment or as a result of conditioning the acetate to a high initial concentration of chloroform) the boundary is no longer visible under the microscope. Its expected position is still made clear by the interferometer because of the change in the fringe spacing. Cellulose acetate conditioned with 26 % of chloroform is an example of this. In polarized light a sharp colour boundary between double refracting and isotropic regions is still visible with this degree of conditioning, and this is in the same position as the fairly sharp bend in the fringe number—distance curve, though no boundary is visible under the ordinary microscope.

No attempt is made in this paper to suggest a numerical value for the limiting value of the maximum $d^2\mu/dx^2$, partly because it is something of a digression from the main purpose of the paper and partly because such a value would presumably not be of universal significance. The limiting value will depend for example on the particular optical system used in the experiment.

In certain cases the middle boundary may indicate, if not a refractive index discontinuity, at least a number of unresolved fringes which appear as a very dark line in the interferometer. This is the case, for instance, when a circle of cellulose acetate is penetrated by methylene chloride as in the experiments described by Hartley (1949); the very pronounced middle boundary as it appears some distance from the 'equator', appears in the interferometer as unresolved fringes. The same sort of line is seen near the corner in figure 7, plate 17.

PROPERTIES ASSOCIATED WITH THE MIDDLE BOUNDARY

We have already remarked that with chloroform the middle boundary separates a region of double refraction from one of optical isotropy. This is not always strictly so. Thus with methylene chloride and with acetone the middle boundary is not

coincident with the boundary of the isotropic region and here the double refraction extends for several fringes into the region on the more dilute side of the middle boundary. This can be seen from photographs of the fringe system taken without polaroid showing the doubling of fringes in the anisotropic region. The significance of this is seen on examining the fringe number—concentration curves. For the chloroform system this curve, as has already been said, consists of two almost linear portions making a considerable angle with one another and joined together by a very short transition region. But in the case of acetone and methylene chloride, this transition region is spread over a considerably greater range of concentration: here the double refraction is found to extend to that concentration of penetrant where the curved transition portion meets the linear portion tangentially, while the highest rate of change in the fringe number—distance gradient (which accounts for the visibility of the middle boundary) is found at a lower penetrant concentration (see figure 11 for example).

In cases where the middle boundary does mark the edge of the isotropic region it marks also the region in which relaxation of molecular orientation occurs. We should therefore expect anisotropy of diffusion to disappear in this region. That this is so can be clearly seen from figure 5, plate 16, where swelling develops parallel to the original direction of stretch once the dilution of the polymer is greater than at the middle boundary. Radial swelling then tends to develop, producing the characteristic curved outer boundaries of figures 5, 6 and 7. It is consistent with this that when the penetrant is one for which the equilibrium swelling is small no middle boundary occurs and consequently anisotropic swelling, in which the corner of the cellulose acetate sheet never loses its original rectangular form, occurs throughout the concentration range. Examples of this are shown in figures 8 and 9 where the penetrants are water and methyl alcohol respectively. For each of the three penetrants for which quantitative data have been obtained (chloroform, acetone, methylene chloride) the middle boundary occurs at approximately the same volume concentration of penetrant. The determination of this concentration is subject to errors arising in the refractive index—concentration relationships, but it appears to lie within 60 to 70 % by volume in each case. The possible significance of this result is referred to later.

From a diffusion point of view the middle boundary is characterized by a relatively sudden increase in diffusion coefficient. An increase is to be expected from the general consideration that the observed relaxation of orientation presumably indicates a general loosening of the polymer structure and greater freedom of movement of the polymer chains. The precise mechanism by which the rate of diffusion is increased is not yet established. It may be that the increased freedom of movement of the polymer results in an increased probability of the occurrence of a 'hole' of suitable size between the polymer chains into which the penetrant molecule can jump. Alternatively, the increased diffusion rate may be due to the fact that the polymer—penetrant gel is less viscous in the relaxed state and hence the mass flow, which accompanies penetration, occurs more readily. Certainly there is a difference in the rheological properties on the two sides of the boundary; for example on the dry side the cellulose acetate sheet containing penetrant can hang

on a hook, but on the dilute side the gel is too fluid and flows under its own weight. Probably both factors contribute something to the increase in the diffusion coefficient at the middle boundary.

The phenomena occurring at the middle boundary bear considerable resemblance to certain second-order transition effects (Boyer & Spencer 1946). A second-order phase transition is characterized by a change in slope of a graph of any of the primary thermodynamic properties against temperature. A marked change in physical properties accompanies this change in slope. Thus for high polymers the second-order transition temperature is one below which any normal rubber-like material loses such properties and above which any dimensionably stable plastic acquires rubber-like characteristics. Similar changes in rheological properties have already been noted as concentration is increased in the neighbourhood of the middle boundary and it is possible that the molecular picture which has been advanced to account for the second-order transition effects may account also for the middle boundary effects. This molecular picture has been summarized by Boyer & Spencer (1946) as follows: 'What appears to happen as the transition temperature is approached from below is that segments of the long polymer chains gradually move further apart, or new "holes" appear. Eventually side groups on the hydrocarbon chains, and whole segments of the long polymer chains, are able to execute more or less free rotations about the long axis of the molecule. The ability of the molecule to act as an articulated chain provides the condition necessary for viscous flow according to the Kauzmann-Eyring (1940) theory of viscous flow in polymers.' If an increase in penetrant concentration is considered to act like an increase of temperature as regards providing more space for segment rotation, this description may apply equally well to the molecular behaviour near the middle boundary. Further evidence in favour of this argument is provided by a molecular model of cellulose acetate. Measurements of the model show that the ratio of the two short axes of the molecule is approximately 3/1. Thus one repeating unit will need about three times its own volume in which to rotate about the long axis of the chain, and the volume fraction of penetrant is therefore about 2/3. This agrees with the figure of 60 to 70 % already quoted as the middle boundary concentration.

It is also possible that some closely bonded or perhaps crystal regions exist in the dry cellulose acetate and that these are mechanically and/or chemically disrupted at the degree of swelling corresponding to the middle boundary. In such a case whereas the polymer vibratory movements occur only in relatively short segments on the dry side of the middle boundary, the whole polymer chains will be free to move on the dilute, relaxed side of the boundary.

The middle boundary also occurs both in unstretched sheet containing some uniplanar orientation and in completely isotropic sheet. With a sheet containing high percentage of plasticizer (the untreated commercial sheet) no middle boundary was visible even in the interferometer.

The occurrence of the boundaries seems to be a phenomenon general to polymer systems which arises with both stretched and unstretched sheet and is not peculiar to cellulose acetate (Hartley 1949). Sharp inner boundaries have been observed with all the polymer and penetrant combinations examined (e.g. water in viscose, polyvinyl

alcohol and gelatin; methylene chloride in polystyrene). Intermediate boundaries between the inner boundary and the outer limit of swelling are also found, provided that maximum swelling is great enough to reach a concentration where relaxation is appreciable. Thus a very marked intermediate boundary is found with gelatin and water. With polystyrene and methylene chloride the 'middle' boundary is more difficult to resolve since complete relaxation occurs at a very low concentration of penetrant. It is therefore close to the inner boundary.

SEQUENCE OF EVENTS AS PENETRANT ENTERS THE POLYMER

The limited data which have so far been obtained do not allow a complete understanding of the nature of the diffusion and swelling but the following may be given as a tentative picture of the successive events.

For diffusion in a direction at right angles to the direction of orientation, it would seem that we have the following successive stages for a penetrant which gives a high degree of swelling. The penetrant will at first be filling the Hermans (1946) free space until the concentration is reached at which this no longer exists. The amount of movement of segments of polymer chains accompanying this stage will vary according to the penetrant used. In an extreme case, as with the very small water molecule, this movement is probably at first negligible. On the other hand with some penetrants, although a refractive index peak is visible, considerable movement of polymer segments may take place simultaneously with the filling of the free space. Until the middle boundary concentration is reached there is no movement, or very little movement, of whole chains parallel to the direction of stretch, swelling taking place almost entirely by the individual movements of a few segments at a time gradually increasing the average distance between the two chains. At the same time there is a gradual decrease in the degree of orientation as a result of these movements. On passing the middle boundary concentration (60 to 70 % by volume of penetrant) the last degrees of orientation are rapidly lost and new movements including rotations of chain segment become possible. When this stage is reached, not only are the chains now able to move parallel to the direction of stretch, but movement of the chains and their segments become equally easy in either direction, as is seen from the swelling now becoming isotropic. This last stage is maintained until the maximum degree of swelling is reached.

In conclusion, the authors wish to thank Miss M. E. Henry for carrying out the calculations, and Mr J. Goodwin for technical assistance with experimental work and photography.

REFERENCES

- Barrer, R. M. 1941 *Diffusion in and through solids*. Cambridge University Press.
Boltzmann, L. 1894 *Ann. Phys., Lpz.*, **53**, 959.
Boyer, R. F. & Spencer, R. S. 1946 Article in *Advances in colloid science*, **2**. New York: Interscience Publishers.
Crank, J. & Henry, M. E. 1949 *Trans. Faraday Soc.* **45**, 1119.
Crank, J. & Park, G. S. 1949 *Trans. Faraday Soc.* **45**, 240.
Darken, L. S. 1948 *Metals Technol. Technical Publication* 2311.

- Hartley, G. S. 1946 *Trans. Faraday Soc.* B, **42**, 6.
Hartley, G. S. 1949 *Trans. Faraday Soc.* **45**, 820.
Hartley, G. S. & Crank, J. 1949 *Trans. Faraday Soc.* **45**, 801.
Hermans, P. H. 1946 *Contribution to Physics of cellulose fibres*. Amsterdam: Elsevier Publishing Co.
Hermans, P. H. 1949*a* *Physics and chemistry of cellulose fibres*. Amsterdam: Elsevier Publishing Co.
Hermans, P. H. 1949*b* Article in *Colloid science*, p. 624, edited by Kruyt, H. R. New York: Elsevier Publishing Co.
Kauzmann, W. & Eyring, H. 1940 *J. Amer. Chem. Soc.* **62**, 3113.
Mark, H. 1943 Article in *Cellulose and cellulose derivatives*, edited by Ott, E. New York: Interscience Publishers.
Matano, C. 1932-3 *Jap. J. Phys.* **8**, 109.
Robinson, C. 1950 *Proc. Roy. Soc. A*, **204**, 339.
-

Statistical kinetics of Hevea rubber cyclization

BY M. GORDON

Dunlop Research Centre, Erdington, Birmingham 24

(Communicated by C. K. Ingold, F.R.S.—Received 21 July 1950)

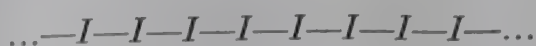
Rubber cyclization provides a useful illustration of statistical effects intervening in the reactions of long-chain polymers. Analogously to the known cyclizations of simple di-isoprenic terpenes, each individual cyclization act links two adjacent isoprene units in the rubber chain into a six-membered ring. However, some of the units remain stranded between neighbouring rings and escape cyclization for want of a partner. Thus the final polymer has a co-polymer structure of 86.5 % of di-isoprenic rings and 13.5 % of uncyclized monoisoprenic units. This result is derived from a statistical theory of Flory, and has previously found accurate experimental verification. The bearing of this statistical effect on the reaction kinetics has now been studied for the first time. Dilatometric experiments on rubber emulsions in sulphuric acid have shown that the cyclization reaction conforms in the main with a statistical cyclization law much better than with a first-order equation uncorrected for the survival of stranded units. High accuracy was required for this experimental distinction, whose importance lies in the implicit proof that the polymer units participate in the rate-determining transition complex, in accordance with the accepted proton transfer mechanism. A kinetic proof has incidentally been furnished for the fact that the rubber polymer contains long chains of identical units. When handled by an emulsion technique, this polymer can provide a useful 'model compound' for its simpler terpene analogues.

Certain rate abnormalities have been observed including an induction period of reduced rate, probably connected with diffusion effects. A small volume change subsequent to the cyclization reaction may reflect a prototropic double-bond shift. The activation energy and entropy are found to be lower than in previous work, an effect traced in part to a curvature of the Arrhenius plot. The known dependence of cyclization rate on Hammett's acidity function H_0 is quantitatively confirmed.

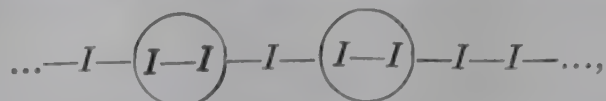
1. INTRODUCTION

Standard chemical reactions can often be carried out, not only on the customary small organic molecules, but equally on regular, high molecular weight polymers. In such cases statistical effects may sometimes be observed to influence both the kinetics and the final composition of the products, whereby light may be shed on reaction mechanism and on structure. In simple cases two identical adjacent

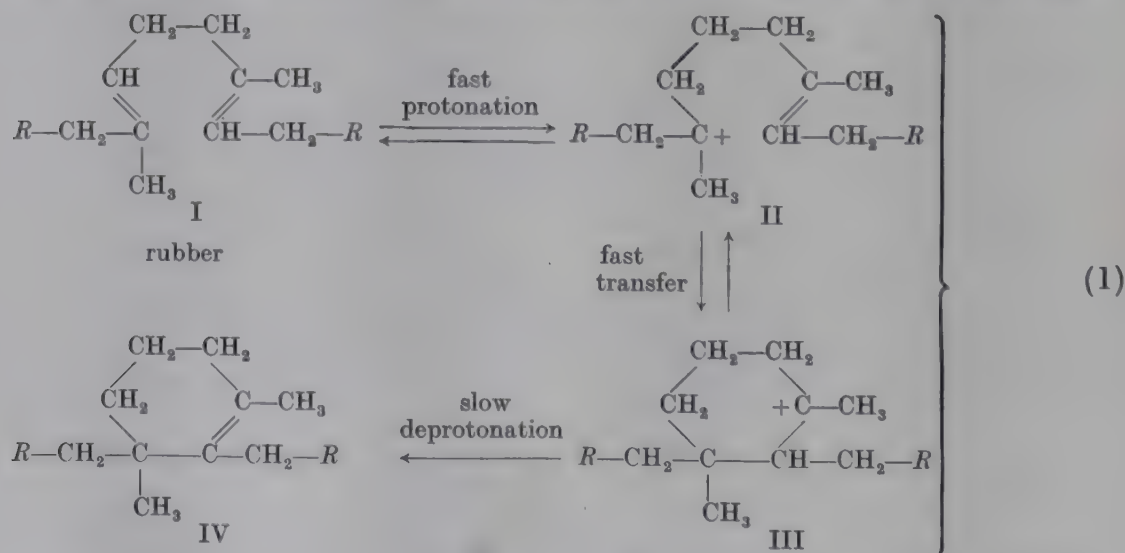
repeating units of a polymeric chain co-operate in an individual irreversible reaction step, usually in order to achieve a ring structure. If the unit be denoted by I , and the polymer chain by I_n or



one may picture the following partly reacted section of a polymer chain:



in which the ringed couples are supposed to have reacted to form rings. Since the isolated I unit lying between the two rings shown is actually precluded from later reaction for want of a partner, the reaction cannot proceed to stoichiometric completion. Flory (1939) calculated the proportion of these 'widows' (isolated units) in the final polymer to be $1/e^2$ (i.e. 13.53 %) of the original I units, on the assumption that at each stage every couple of adjacent unreacted units has an equal chance of reaction. This result is practically independent of the degree n of polymerization, provided n exceeds about 15 or 20. Flory (1939) and Wall (1940) extended the calculation to the more complicated reactions of co-polymers and obtained results of importance for the experimental investigation of the arrangement of monomer units in the chain (Marvel & Levesque 1939; Gordon 1949, 1950) pointed out that the cyclization of rubber is subject to this statistical effect and that analytical data on the unsaturation percentage of cyclized rubber by Fisher & McCollm (1927), which were previously unintelligible, actually provide the best known experimental verification of Flory's theory. The kinetics of this cyclization, catalyzed by 70 to 80 % sulphuric acid, were studied (Gordon 1949, 1950) by an emulsion technique on the basis of measurements in a diffusion gradient tube of the density of purified samples of the flocculated polymer. The mechanism supported by this work was:



Two of the original isoprenic units I (viz. $\begin{array}{c} \text{CH}_3 \\ | \\ -\text{CH}_2-\text{C}=\text{CH}-\text{CH}_2- \end{array}$) in rubber here join to form a ring such as IV, and a theory was deduced for the statistical kinetic effect. This is due, of course, to the variations in status of an I unit as regards its

chance of reacting, according to the number of its neighbours (0, 1, or 2) already reacted at time t during the reaction. The effect should manifest itself in a deviation from unity of the reaction order, initially to 1.5, and at the end of the reaction to about 1.16. Such deviations were not detectable by the gradient tube technique, and the present work was undertaken to confirm the rate equation by a more sensitive dilatometric method.

The chief value of the accurate confirmation actually obtained below lies in the consequent proof that the polymer forms part of the activated complex which determines the rate of the reaction. Such complexes involving polymers (cf. the propagation step of polymerization) are interesting as regards the entropy of activation and in other ways. Many simple di-isoprenic terpenes such as dihydro-myrcene cyclize under similar conditions as rubber, and undoubtedly by an analogous mechanism. Since they are not subject to the statistical effect under discussion, however, they do not lend themselves to this direct demonstration of their participation in the rate-controlling process. In fact, kinetic measurements on these terpene cyclizations are still not available, though the emulsion technique should render them quite feasible (Gordon 1950), and the polymer rubber has proved to be a suitable 'model' compound for the lower terpenes, not only because of its advantageous statistical effect, but for its ease of manipulation also. Apart from throwing light on the role of the polymer in the mechanism, the measurements here reported strengthen the conclusions concerning the part played by the sulphuric acid. The fair correlation between reaction rate and Hammett's acidity function previously noted is quantitatively confirmed. The conclusion that deprotonation rather than protonation is rate-determining still appears to hold, though one of its supports is weakened by a reduction in the accepted frequency factor resulting from these more precise measurements.

2. DERIVATION OF THE STATISTICAL KINETIC EQUATION

It will be assumed that cyclization is initiated by suitable activation on an isoprene unit I in the rubber chain, followed by attack of this unit on its right-hand neighbour to form a ring (e.g. as in (1)), provided this neighbour has not already cyclized. The relative rates of these two steps, and the degree of reversibility of the first step (activation), affect only the meaning of the overall rate constant k_s of the statistical kinetic equation deduced below. Nor is the equation affected if activated units are allowed to cyclize with their left as well as their right neighbours at rates remaining proportional to each other. The law of mass action in terms of weight concentrations is assumed, and the derivation may be visualized as applied to one rubber chain with N (Avogadro's number) units, weighing 68 g.

Imagine the chain cut at time t into chain segments and ring segments by severing all junctions between rings and uncyclized units, or between one ring and another. Denote the weight fraction of ring segments by r , this being a convenient measure of the degree of cyclization attained. The weight fraction of chain segments of exactly k adjacent I units may be denoted by i_k . Thus

$$r + \sum_1^{\infty} i_k = 1 \quad \text{and} \quad dr/dt = - \sum_1^{\infty} di_k/dt. \quad (2)$$

The exact rate equation for the growth of r is

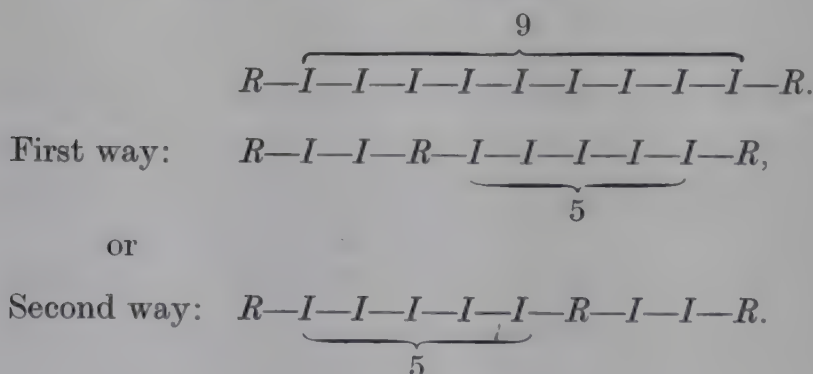
$$dr/dt = k_s \sum_1^{\infty} i_k (k-1)/k = k_s (1-r - \sum_1^{\infty} i_k/k) = k_s (1-r - (N_c/N)). \quad (3)$$

This is deduced by noting that rings are formed from the various chain segments comprising k adjacent I 's at rates proportional to i_k (law of mass action), but also proportional to $(k-1)/k$, because each segment contains one out of its k units which is ineffective. The ineffective unit is the one at the right end, whose activation cannot lead to cyclization for want of a right-hand partner. In the second equality of (3), the term $\sum_1^{\infty} i_k/k$ represents the deviation from first-order kinetics, and in the third equality this is equated to N_c/N , where N_c is the total number of chain segments irrespective of length.

The weight fraction i_k has a rate of change which is the balance of its rate of formation and rate of decay:

$$di_k/dt = k_s k \left(\underbrace{2 \sum_{k+2}^{\infty} i_n/n}_{\text{formation}} - \underbrace{i_k (k-1)/k}_{\text{decay}} \right). \quad (4)$$

Thus its rate of change is proportional to the number k of I units involved. The term $k+2$ under the summation sign indicates that the chain segment with k units can be created only from segments having at least two extra units. The multiplier $2/n$ affecting i_n applies, because activation of two only out of the n units in an n -chain is favourable to such a creation. For instance, a segment of five can be created out of one of nine units in the following two ways:



The term in (4) for the rate of decay is explained as the similar terms in (3). One can simplify the system of equations (3) and (4) by eliminating all fractions except r , i_1 and i_2 :

$$dr/dt = k_s (1-r-i_1-i_2/2 - di_1/dt). \quad (5)$$

Further simplification, leading to the exact functional dependence of r alone on t , though theoretically possible, seems difficult. A very satisfactory approximate solution will now be derived by estimating the term N_c/N in (3). This solution will be integrable, will agree with the exact solution asymptotically at the beginning ($r=0$) and end ($r_f = 0.8647 = 1 - 1/e^2 = 1 - i_f$), and will rest intermediately on a plausible physical picture of the approximate distribution of the segments.

If N_r is the total number of rings at any time t , we have

$$N_r/N = \frac{1}{2}r. \quad (6)$$

Let X be the fraction of the rings which are neighboured on the right by another ring, i.e. a parameter measuring the degree of 'clustering' together of the rings in the distribution of rings and chains during cyclization, so that

$$N_c/N = (1 - X) N_r/N = \frac{1}{2}(1 - X)r, \quad (7)$$

since the number N_c of chains is equal to the number of those rings which are *not* neighboured by a ring on the right. The rate equation (3) becomes

$$dr/dt = k_s(1 - r - \frac{1}{2}(1 - X)r). \quad (8)$$

At the beginning of cyclization, when r is very small, X must approach zero, while later as r grows, X will grow. Putting X equal to zero, the initial rate equation becomes asymptotically

$$dr/dt = k_s(1 - 1.5r), \quad \text{initial reaction (1.5 order)}. \quad (9)$$

This law does not represent the further course of the reaction, e.g. it brings cyclization to an end ($dr/dt = 0$) when $r = 0.6667$ instead of 0.8647. To find X approximately as an increasing function of r , one notes that the isoprenic unit lying just to the right of a ring may either be found cyclized with its right neighbour or not, and that X is the ratio of the probabilities of these two events. If we choose an isoprene unit entirely at random (without requiring it to lie to the right of a ring), the probability of it being found cyclized with its right-hand neighbour—the same as that of being found cyclized with its left neighbour—is $\frac{1}{2}r$. The probability of it *not* having cyclized with its right neighbour is $1 - (\frac{1}{2}r)$, and the ratio of the two events of interest is accordingly $r/(2 - r)$ for an entirely arbitrarily chosen unit. Assuming that this represents also the required ratio for a unit lying just to the right of a ring, an assumption probably near to the truth, there results

$$X = r/(2 - r). \quad (10)$$

One can test this assumption at the end of the reaction by substituting (10) in (8) and putting dr/dt equal to zero. Solving for r , one obtains $r = 1$ instead of 0.8647 as the final degree of cyclization. It is clear that the 'clustering' tendency X of the rings at the end of the reaction has been somewhat overestimated by (10), but a mere 10 % adjustment in X is required, as shown by the term 0.902 in (11), to obtain the correct asymptotic behaviour ($r_f = 0.8647$) in (12):

$$X = (r/(2 - r)) \times (e^2 - 3)(e^2 + 1)/(e^2 - 1) = 0.902r/(2 - r), \quad (11)$$

and by substitution in (8)

$$dr/dt = k_s(1 - r - (\frac{1}{2}r)(1 - 0.902r/(2 - r))). \quad (12)$$

On the basis of the underlying physical meaning, there is little reason to suppose that X will deviate at all seriously from (11) in the earlier parts of the reaction, where in any case the equation (12) becomes progressively less sensitive to such

deviations. Further adjustments are accordingly not warranted, and the desired approximate rate equation (12) is integrated to

$$t = F(r) - F(r_i) = k_s \{ -0.3569 \log (0.8674 - r) / (1.187 - r) - 0.5909 \log ((1.0257 - r)^2 - 0.259) - 0.4846 \} - F(r_i), \quad (13)$$

where r_i in the integration constant $F(r_i)$ denotes the initial degree r of cyclization. It is shown in (5) that experimental data confirm this equation in preference to a first-order law.

3. THE EQUILIBRIUM POLYMER AND ITS STRUCTURE

The statistical kinetic theory of equations (3) and (4) and the statistical thermodynamic theory of Flory (1939) can be unified and checked against each other. The following infinite series is postulated:

$$di_1/dt = - \sum_3^{\infty} (S_n/n) (di_n/dt), \quad (14)$$

and permissible values of the coefficients S_n found by comparing coefficients of i_n in equations (4), using successively the values $k = 1, 2, 3$, etc. The result is

$$S_n = 2(1 + S_3 + S_4 + \dots + S_{n-2})/(n-1), \quad (15)$$

which is precisely Flory's equation (3) for the number S_n of units I out of an n -chain which will on average be 'widowed' in the whole course of the reaction. Integration of (13) yields

$$i_1 = i_f - \sum_3^{\infty} S_n i_n / n, \quad (16)$$

which correctly asserts that the current fraction i_1 of widows at time t is equal to the final value minus the fraction yet to be widowed from among all the chain segments remaining at that time. The distribution function of rings and chain segments during, or of R and I units at the end of the reaction, though implicit in equations (3) and (4), has not been solved. A reasonable *model* with a calculable distribution function for the statistically cyclized rubber can be derived. It correctly contains 13.53 % I units and 86.47 % R 's by weight, i.e. about three rings to an isoprene unit, in a random co-polymer structure schematized thus



As cyclization is irreversible, two I 's must not come together. The model is constructed by imagining $3 \times 13.53 = 40.59$ % of a monomer IR , consisting of an isoprenic unit *flanked on its right by a ring*, to co-polymerize randomly with 59.41 % of a monomer R . The distribution function of this model is readily calculated, and from it the interesting conclusion can be drawn that 51.54 % by weight of the unit RIR can be cut from the cyclized rubber by judicious splitting of links in the polymer chain (17). This shows the usual representation of cyclized rubber as $(R)_n$ to be quite inadequate.

4. EXPERIMENTAL

About 1300 g. of acid latex were compounded, essentially as described in previous reports (see also B.P. 634,879), by mixing Dunlop 60 % Hevea Rubber Latex with a stabilizer and sulphuric acid. Four kinetic runs were undertaken over a period of 4 months by following the volume change of the latex during cyclization in the thermostatically controlled Pyrex dilatometer shown schematically in figure 1. At elevated temperature the stem of the dilatometer passed through a lid covering the thermostat (not shown), and the control was to one-thirtieth degree or better. The fifth run recorded in table 1 was intended as a check, being based on a latex compounded from a different shipment, and by its excellent concordance proves that the rates are reproducible. The run numbered 4 was on a latex obtained from the acid latex of runs 1 to 3 by dilution with a known amount of water, to check the influence of acidity on rate (see § 6). The full details of measurement and checks quoted hereafter refer to the model run 3, comprising 66 measured points taken consecutively during one day. Detailed calculations were also performed on runs 2 and 5, and gave on all accounts substantially equally good agreement. In particular, the results fitted the statistical law better than a first-order equation.

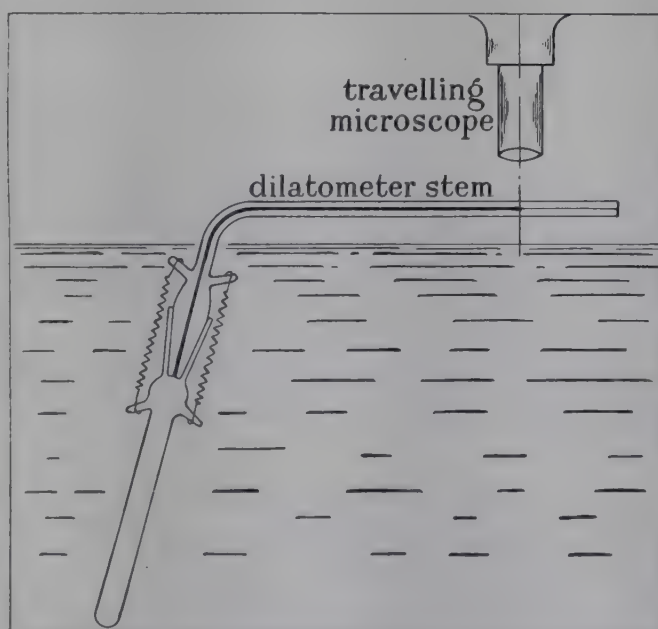


FIGURE 1. Sketch of dilatometer.

TABLE 1

run no.	temp. (° K)	percentage concentration <i>c</i> of H ₂ SO ₄	rate constant <i>k</i> ₀ (min. ⁻¹)
1	298.5	78.1	0.00021
2	332.7 ₇	78.1	0.01732
3	338.5	78.1	0.02652
4	357.2 ₆	70.9	0.013
5	298.5	79.2 ₃	0.0002985
5	298.5	corrected to 78.1	0.000192

The acid latex was stored in a dropping funnel at -20°C in a frozen condition. When required for filling the dilatometer, it was allowed to attain about $+10^{\circ}\text{C}$, stirred and evacuated to 1 in. of Hg for about 10 min. The evacuation was found essential to avoid the evolution of a small amount of gas during the dilatometric run. The dilatometer was filled and weighed, thus giving the initial density D_i^L of the latex. Simultaneously, a sample was drawn for titration of the acid content by alkali, and one flocculated for the determination in a diffusion gradient tube of the density D_0^P of the pure polymer phase at that instant. The weight percentage of rubber in the latex was slightly variable from run to run owing to creaming effects of the heavy sulphuric acid phase on the light rubber phase. The effect of this on the overall composition of the large remaining latex bulk was negligible. The correct rubber content of the latex in the dilatometer was calculated in two different ways, on the assumption that creaming did not disturb the kinetically significant concentration c ($\text{H}_2\text{SO}_4/(\text{H}_2\text{SO}_4 + \text{H}_2\text{O})$) of the acid. The first calculation was based on D_0^L and D_0^P , the second on the acid titration. Close agreement between the two values obtained justified the assumption just mentioned, and the value based on the titration was accepted. The dilatometer was immersed in the thermostat and 6 min. allowed for thermal equilibration. The initial zero reading of the latex meniscus in the stem was then taken. The density D_i^P of the polymer at this instant was still close to D_0^P , the small change during equilibration being readily estimated from the subsequent rate. The isothermal contraction during the run was followed by noting the meniscus displacement M (cm.) from its original zero position to its final M_f . The total volume contraction during cyclization being less than 2%, no stem correction was necessary for the cold latex drawn from the dilatometer stem into the dilatometer during the run. The omission of this very small and uncertain correction favours a first-order law over the statistical one, since it must cause a drop in the apparent reaction order. Curves showing the reaction progress r against time t are plotted in figure 2. For this purpose r was calculated from the linear relation (21) linking it with M . Table 3 gives all the observed values of M and t for run 3.

Check on the validity of the technique

The rate constants of the dilatometric runs 2, 3 and 4 agree within a factor of two to three with the previous work based on measurement of D^P in a gradient tube. This would be accounted for by an error of about 2% in the sulphuric acid concentration c , but agreement is actually somewhat better, because the present rate constants k_s refer to the statistical law and take cognizance of the induction period (see § 7), while the previous work employed first-order constants k_1 . The much less satisfactory agreement of the k_s values for the runs at 25.3°C with the previous rate constants reflects a lower activation energy than previously found, a discrepancy discussed in § 7.

As an independent check on the validity of the present method the total measured volume contraction was calculated from M_f and compared with the separately measured density change of the polymer phase proper. To this end the practically stationary final density D_f^P of the polymer was measured in a gradient tube at the end of the run, and D_i^P calculated from M_f for comparison with the measured value

(i.e. the corrected value based on D_0^P). As shown in table 2, D_i^P calculated from M_f according to formula (18) was 0.920_4 , while D_0^P was found 0.920 ± 0.001 and D_i^P estimated therefrom 0.922 ± 0.002 . This agreement shows satisfactorily that the dilatometric volume change of the latex reflects the volume change of the rubber particles in it during cyclization, especially if it is borne in mind that small effects due to non-rubber constituents in the latex have been neglected and that a large number of measurements enter into the calculation. In particular, the dilatometric volume change being measured at 65.3°C , but the polymer densities at 25.3°C ($\Delta T = 40^\circ \text{C}$), a correction was required for the difference of the apparent coefficients

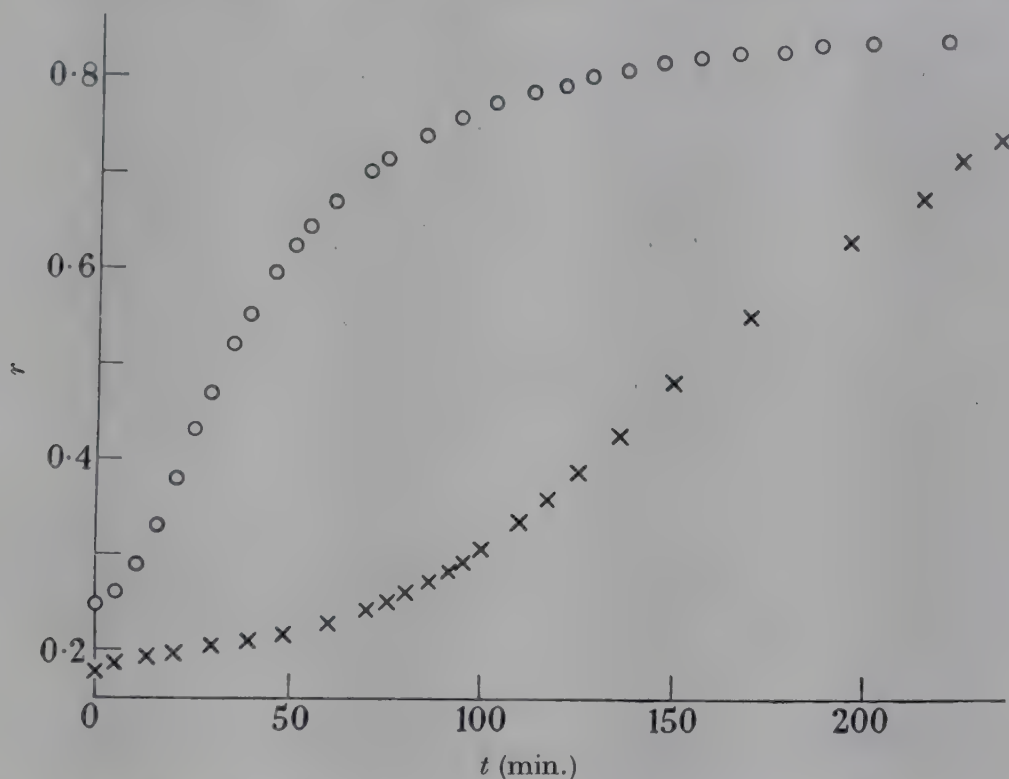


FIGURE 2. Plot of degree of cyclization r against t min. O, run 3: x, run 4.

of expansion of the initial latex C_i^L and the final latex C_f^L . Thus since the coefficient of expansion of the starting polymer (slightly cyclized rubber) was about 0.0005, but that of the fully cyclized rubber about 0.0002 only, the volume contraction of the latex at 65°C was about 16 % larger than at 25°C . The C^L values were measured in the same dilatometer over the same range T , so that disturbing effects such as the small expansion of the Pyrex dilatometer itself could be neglected. All the measurements are recorded in table 2 and the formula employed was

$$1/D_i^P = 1/D_f^P - \frac{M_f - \Delta T V dM/dV (C_i^L - C_f^L)}{m(dM/dV)}. \quad (18)$$

On the assumption that the deprotonation step in (1) is rate determining, an upper limit previously estimated for the basic strength of rubber (pK_a of the carbonium ion III < about -8) is here experimentally confirmed. Thus since the volume change during cyclization must be associated very largely with the proton transfer step which precedes the slow deprotonation, the protonation and transfer equilibria in

(1) cannot lie more than a few per cent over to the products, otherwise the slow contraction observed on the latex would correspondingly fall short of the contraction of the rubber phase during cyclization.

TABLE 2. DETAILS OF MEASUREMENTS FOR RUN 3

immersed volume of dilatometer	V	$= 3.754_2$
rate of meniscus movement with latex volume	dM/dV^L	$= 128.7 \text{ cm.}^{-2}$
initial density of latex at 25° C	D_i^L	$= 1.4985 \text{ g./ml.}$
rubber content calculated from compounding, by weight		19.1%
actual accepted rubber content (see text)		16.3%
mean mass of rubber in V during run (calc. at 25° C)	m	$= 0.926 \text{ g.}$
observed final meniscus displacement	M_f	$= 9.45 \text{ cm.}$
density of (slightly cyclized) starting polymer ^a (25° C)	D_0^P	$= 0.920 \pm 0.001$
estimated polymer density at first meniscus reading (25° C)	D_i^P	$= 0.922 \pm 0.002$
D_i^P calculated from formula (18) as a check		$= 0.920_4$
final density of cyclized polymer (25° C)	D_f^P	$= 0.981_3$
mean apparent coefficients of cubical expansion measured:		
H_2SO_4 (79%)		$= 0.00055_1 \text{ }^\circ\text{C}^{-1}$
initial latex (slightly cyclized)	C_i^L	$= 0.00053_7 \text{ }^\circ\text{C}^{-1}$
final latex (fully cyclized)	C_f^L	$= 0.00046 \text{ }^\circ\text{C}^{-1}$

5. DISCRIMINATION BETWEEN STATISTICAL KINETIC AND FIRST-ORDER LAW

It will now be shown that the main central portion of run 3, covering about 78 % of the measured volume change and 50 % of the total cyclization reaction, conforms definitely to the statistical equation (19), with $F(r)$ as defined in (13), rather than to the first-order law (20). The comparison is exhibited in table 3 by confronting the observed times $t_{\text{obs.}}$ with times t_s calculated from (19) and t_1 calculated from (20):

$$k_s t_s = F(r) - F(r_i) + x, \quad (19)$$

where $r_i = 0.2468$, $F(r_i) = 0.3052$, $x = 0.3052$, $k_s = 0.02652 \text{ min.}^{-1}$;

$$k_1 t_1 = -\log(1-r) + \log(1-r_i) + x', \quad (20)$$

where

$$r_i = 0.1879, \quad -\log(1-r_i) = 0.7261, \quad x' = 0.7368, \quad k_1 = 0.010307 \text{ min.}^{-1}.$$

The calculation of the times and of the r 's are respectively slightly complicated because a small initial portion of all the runs, and a very small final portion of the runs at elevated temperature, conformed with neither law. Thus in fitting the experimental $t_{\text{obs.}}$ and the r values calculated from the measured meniscus displacements M to equations (19) and (20), ideally only the rate constants k_s and k_1 require adjustment. In practice, however, two *small* additional adjustments have to be made in both cases to allow for the initial and final deviations. These adjustments do not conceal any favoured treatment of either law. Thus the first adjustment concerns a displacement of the time origin by an amount x or x' respectively as shown in the last two equations, to allow for an 'induction period', which is visible in figure 2 and whose significance is discussed in § 7. Probably fortuitously, these adjustments cancelled the integration constants $F(r_i)$ and $\log(1-r_i)$ in runs two and three within experimental error, so that the main part of the reaction proceeded along the theoretical curve for $r_i = 0$ and absence of an induction period.

The meniscus movement reflecting accurately (§ 4) the volume change in the rubber phase, is therefore linearly related to the reaction progress:

$$aM = r - r_i, \quad (21)$$

where for equation (19), $a = 0.0669$, $r_i = 0.2468$; and for equation (20), $a = 0.0897$, $r_i = 0.1879$. Ideally, the constant a should be fixed as $(r_f - r_i)/M_f$. However, the abnormality at the end of the reaction (§ 7) causes a slow volume change after (on the kinetic evidence) cyclization is complete, and thereby affects M_f and a . For equation (19), $r_f = 0.8647$ and $(r_f - r_i)/M_f = 0.0654$, while 0.0669 gives the best fit. Again, for equation (20) $r_f = 1$, the theoretical $a = 0.08593$, while $a = 0.0897$ gives the best agreement. The r_i values chosen correspond closely to those calculated from D_i^F and the density (0.900) of uncyclized rubber. It is emphasized that after suitable adjustments in a and k_s (but not k_1), variations in the assumed r_i of ± 0.1 leave the fit of equation (19) practically, and that of equation (20) exactly, unaltered, because of an invariance property of the first order law. Table 3 lists the times $t_{\text{obs.}}$, and the times t_s and t_1 , respectively calculated from equations (19) and (20), with the aid of (21).

After an induction period of about 20 min., 28 out of 49 consecutive values of t_s , covering an interval of 180 min., lie within 30 sec. of the experimental $t_{\text{obs.}}$, 12 points lie between $\frac{1}{2}$ and 1 min., 8 points between 1 and 2 min., and only one point between 2 and 3 min. off. On the other hand, the t_1 values, while agreeing about equally well over the range from 20 to 70 min., and again at 200 min., deviate between 70 and 200 min. smoothly up to a maximum of over 12 min. from the $t_{\text{obs.}}$. These facts show not only that a conclusive experimental decision in favour of the statistical law (19) has been obtained, but also the high accuracy actually required for such a decision.

6. DEPENDENCE OF RATE ON ACIDITY AND TEMPERATURE

The usual Arrhenius plot of the runs listed in table 1 is shown in figure 3. The best straight line through runs 1, 2, 3 and 5 has been drawn, and from this the rate at 357.26°K and sulphuric acid concentration $c = 78.1\%$ can be computed. When this is compared with run 4, at the same temperature but at $c = 70.9\%$, the relation

$$\frac{d \log k_s}{dc} = 0.165 \quad (22)$$

is found between these two concentrations, in very good agreement with the value (0.17) obtained in the previous work, so that the correlation between reaction rate and Hammett's acidity function H_0 (Gordon 1950) is confirmed. The small correction of the rate of run 5, from the measured value at $c = 79.23$ to the desired value at $c = 78.1$, as shown in table 1, is based on equation (22).

The temperature dependence of the rate of cyclization agrees much less well with that previously reported, the activation energy calculated from runs 1, 2, 3 and 5 being 25.3 kcal. instead of 32.7, and the frequency factor $4 \times 10^{14} \text{ min.}^{-1}$ instead of 3×10^{19} . The main reason for these deviations must be sought in the runs at 25.3°C , which showed much higher rates than predicted from the previous work. Two

TABLE 3. DISCRIMINATION BETWEEN FIRST-ORDER AND STATISTICAL
CYCLIZATION KINETICS

point no.	meniscus displacement M (cm.)	time (min.)		
		$t_{\text{obs.}}$	$t_{\text{s calc.}}$	$t_{\text{1 calc.}}$
0	0.0	0	11.51	9.81
1	0.148	4	12.12	10.51
2	0.220	5	12.39	10.85
3	0.315	6	12.78	11.31
4	0.562	9	13.81	12.51
5	0.655	10	14.19	12.97
6	0.761	11	14.68	13.51
7	1.002	13	15.74	14.75
8	1.134	14	16.33	15.45
9	1.270	15	16.97	16.17
10	1.407	16	17.62	16.92
11	1.714	18	19.16	18.65
12	1.865	19	19.95	19.53
13	2.020	20	20.76	20.44
14	2.324	22	22.45	22.31
15	2.588	24	23.98	24.00
16	2.777	25	25.14	25.25
17	3.064	27	26.97	27.22
18	3.195	28	27.84	28.15
19	3.338	29	28.84	29.38
20	3.871	33	32.80	33.31
21	4.126	35	34.87	35.44
22	4.344	37	36.77	37.36
23	4.571	39	38.84	39.43
24	4.765	41	40.75	41.29
25	4.949	43	42.63	43.14
26	5.226	46	45.65	46.09
27	5.407	48	47.79	48.13
28	5.649	51	50.87	51.02
29	5.806	53	53.00	53.02
30	5.968	55	55.33	55.16
31	6.166	58	58.39	57.96
32	6.351	61	61.46	60.76
33	6.541	64	64.87	63.81
34	6.706	67	68.06	66.68
35	6.841	70	70.94	69.17
36	7.018	74	74.93	72.68
37	7.177	78	78.87	76.12
38	7.390	84	84.80	81.20
39	7.675	93	94.11	89.10
40	7.777	97	97.91	92.36
41	7.900	102	102.93	96.62
42	8.018	107	108.29	101.17
43	8.096	112	112.15	104.47
44	8.192	117	117.32	108.91
45	8.229	120	119.47	110.75
46	8.302	124	124.02	114.68
47	8.347	127	127.00	117.25
48	8.403	131	130.91	120.73
49	8.464	136	135.54	124.89
50	8.543	141	142.24	130.96
51	8.558	145	143.54	132.27 !
52	8.613	149	148.89	137.20
53	8.681	155	156.21	144.27
54	8.721	159	160.84	149.08
55	8.723	162	161.22	149.35
56	8.761	165	165.98	154.50
57	8.772	168	167.39	155.98
58	8.804	172	173.44	161.10
59	8.838	177	177.43	167.38
60	8.879	187	184.36	176.08
61	8.916	192	191.45	186.36
62	8.954	200	199.58	200.00
63	8.962	205	201.73	203.45
64	9.007	220	213.49	231.63
65	9.349	316	∞	∞
66	9.450	393	∞	∞

explanations may be advanced for these deviations. First, while the overall activation energy of 32.7 kcal. previously found referred to a temperature range between 25 and 100° C, the Arrhenius plots showed some definite curvature, and lower values can be calculated from the data falling within the range 25 to 65° C here studied. Thus the three points in the earlier work at 78.5 % acid in this temperature range correspond to 29.8 kcal., the two points at 72.6% acid to 24.8 kcal. of activation energy. Secondly, the previous density technique was based on the assumption of a unimolecular law and was insufficiently sensitive to show the initial and final rate abnormalities observed in this work (§ 7), and is accordingly regarded as less reliable in the calculations of the Arrhenius parameters of the cyclization step.

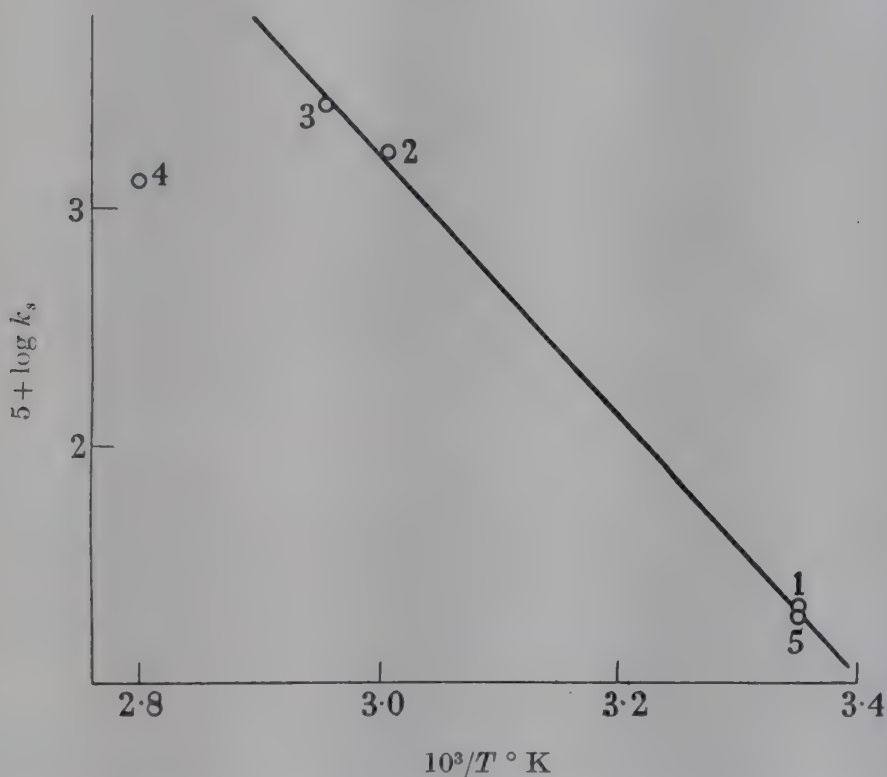


FIGURE 3. Arrhenius plot for runs 1 to 5. (For numbering cf. table 1.)

The only serious implication of the lowering of the Arrhenius parameters at 78.1 % acid and between 25 and 65° C which has thus been accepted is a weakening of the support previously derived from the exalted frequency factor for the rate control by the deprotonation step (as in scheme (1)) rather than by the first (protonation) step. It has been shown (Gordon 1949, 1950) that a very high-frequency factor such as 10^{19} min.⁻¹ could correlate only with a reaction step in which ionic charges were neutralized in the transition state, e.g. a collision between the cyclo-carbonium ion III and an HSO_4^- ion to produce IV and H_2SO_4 . The new value of 4×10^{14} is now normal for a unimolecular reaction, though still very high for a collision between an ion and a large uncharged molecule, such as occurs in the protonation step. Moreover, from the previous experience cited, the frequency factor may be expected to rise with temperature, and certainly with acid concentration c , so that considerably higher values could actually be experimentally realized. Since the deprotonation rate control is by no means an unusual mechanistic

feature, and has been supported also by the influence of the structure of various terpene substrates on their approximately known cyclization rates, it is still the preferred mechanism though on weaker grounds in view of the reduced activation entropy.

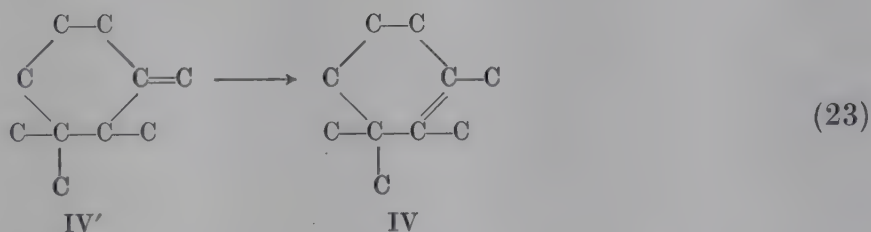
7. SMALL INITIAL AND FINAL RATE ABNORMALITIES

An induction period of reduced rate was observed in the runs at elevated temperature. While, as figure 1 shows, this was particularly marked at the reduced acidity of run 4, it did not interfere even there with the determination of the rate constant of the later regular part of the reaction. Trials showed that the detailed course of this induction period was accurately reproducible for a given latex. As at 25° C the initial rate abnormality was much reduced in importance relative to the main change, it follows that whatever the mechanism responsible for it, it must be subject to a much smaller activation energy. The abnormal rate period split into two phases at 25° C. The first phase on bringing the latex from a low temperature to 25° C and equilibrating, was a 'burst' of volume contraction, lasting for about 100 min., at first several times too fast compared with the later regular reaction. This slowly died down to a rate below the normal value and then rose back to the normal in a few hours. These effects were likewise reproducible and confined to the first 10 to 20 % of the total measured contraction. Whether the 'burst' occurred also at elevated temperature is uncertain, as it would there be obscured by the time of equilibration of temperature, though it is visible in the slow run (figure 1). Its significance is not understood, and it may indeed be connected with some side reaction in the complex colloidal latex system, rather than with cyclization. The induction period of reduced rate may be connected with the rate of diffusion of the acid catalyst into the rubber particles (Gordon 1949), a conclusion supported by its low activation energy.

Detailed consideration shows that the small induction period cannot reasonably be accounted for by any chain reaction mechanism for cyclization. The most plausible chain mechanism would have the effect of cyclization of a couple of isoprene units spread to an adjacent couple, and so on all along the segment of chain remaining uncyclized. This mechanism would arise, for instance, if the deprotonation step of one ring became the protonation step of the adjacent unit. Such a scheme would bear only superficial resemblance to the acid-catalyzed chain polymerization of olefins, though a number of radical reactions of rubber have been formulated as similar chain sequences of cyclizations involving two units at a time (see § 8). The correlation of cyclization rate with Hammett's acidity function H_0 indicates a proton transfer from the acid to the rubber; this step would actually have to be rate controlling in any chain mechanism, as the alternative of deprotonation control (as in (1)) would not be available. A chain mechanism controlled by the activating protonation would, however, have a higher reaction order than here observed, indeed as high as two under the simplest assumptions. This will be appreciated from the simple argument that at half-cyclization ($r = 0.5$), the rate of cyclization would, compared with the starting rate, be halved because the number of units available for activation would be halved, and then again be halved because each activation

would, on average, lead to the rapid cyclization of chains of only half the original length. Thus it does not appear possible that any chain mechanism would produce rate equations that could approach the excellent agreement with experiment achieved by the non-chain mechanism (1) and equation (19). Moreover, the number of widowed units in a chain mechanism would fall to a much smaller proportion than the 13.5 % of Flory's theory for a non-chain mechanism, which is in close agreement with Fisher & McCollm's analytical results on cyclized rubber.

Turning now to the rate abnormalities at the end of the reaction, these reflect a step subsequent to cyclization proper, since the density of the rubber increases gradually well above the end-point approached kinetically during the main part of the reaction. At the highest temperatures previously employed (90 to 100° C), at which there is some danger of oxidative side reactions, densities up to 0.990 had been recorded. It has since been found that at $c = 78\%$ and 60° C ten different latexes gave definite end-products whose densities lay between 0.984 and 0.985, even after several hours continued maintenance under these reaction conditions. The reproducibility of this 'cyclized rubber' must be regarded as high, since most synthetic polymers are sold commercially on density specifications to the second decimal place only. The polymer density approached by the main part of the reaction in the high-temperature dilatometer runs was near to 0.980. The nature of the subsequent shift to 0.984–5, which took several hours to complete, might be sought in a cross-linking or in a prototropic displacement of the remaining double bonds. The latter alternative seems more likely, since there appears to be no marked decrease in solubility or swellability. There is some intrinsic probability that the thermodynamically less stable γ -terpene form IV' is formed first, and the subsequent reaction might thus be its transformation to the β -form IV:



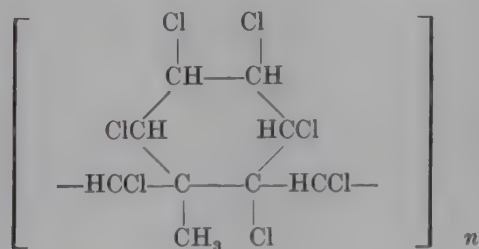
Some spectroscopic evidence has been quoted earlier (Gordon 1949) for the presence of rings IV' in cyclized rubber, though a detailed confirmation of the reaction (23) will require further spectroscopic examinations to be made. (23) would account, not only for the small volume decrease found, but also for the marked rise in softening point, because the semicyclic double bond would confer mobility on the ring IV', while the endocyclic bond in IV would cause considerable stiffening.

In the runs at 25° C the reaction was too slow to carry the polymer in practice beyond a density of 0.974.

8. CONCLUSION

The emulsion technique of making kinetic measurements on latex rubber, presents, among other advantages, much higher precision than earlier methods. The average reaction order of the present measurements being approximately 1.25, the conclusive

discrimination from a 1.0 order law necessitated not only the highest attainable precision, but also a careful investigation (§ 5) of the mathematical effects of possible errors in all the parameters involved. As an illustration of the difficulties of the earlier methods, the simple unaccelerated vulcanization of rubber with sulphur may be cited. Several investigators, working with bulk (as distinct from latex) rubber, assessed the order of this reaction at values ranging from zero to over 1. The correct value is undoubtedly unity, as shown by the analysis of adiabatic rate measurements (Gordon 1948), in agreement with the most likely mechanism. The rate-determining step being the splitting of a sulphur ring without aid from the polymer, statistical *kinetic* effects are not encountered in vulcanization. Numerous examples of statistical kinetic and equilibrium effects probably do exist in rubber chemistry. Farmer (1946) has described several reactions of other substances (SO_2 , Cl_2 , maleic anhydride) with two neighbouring units of rubber at a time. Statistical effects should apply to all these at equilibrium through the survival of uncyclized units; and wherever, in the absence of chain mechanisms, rubber enters the rate-controlling step, equation (13) may be expected to apply. As an example, the empirical formula $(\text{C}_{10}\text{H}_{11}\text{Cl}_7)_n$ has been accepted by Le Bras & Delalande (1950) in support of the structure



which is closely analogous to IV. Uncyclized units, probably $\begin{array}{c} \text{HCCl}-\text{CCl}-\text{CHCl}-\text{CHCl}, \\ | \\ \text{CH}_3 \end{array}$

should accompany these rings, in amounts varying from 13.5 % down to negligible proportions, as the kinetic chain varies in length from units (single cyclization act without lateral propagation) to large numbers. The diagnostic value in studies of reaction mechanism of the appropriate statistical analysis is thus evident, though the experimentally attainable precision is here still the limiting factor.

The present work has confirmed the postulated mechanism by its demonstration that the rubber polymer enters the rate determining step in cyclization. It has established that this reaction is essentially not an intermolecular cross-linking, but a reaction between couples of neighbouring units in a long chain, 13.5 % of which escape from cyclization. Indeed, a purely kinetic proof has now been furnished of the basic fact that rubber is a polymer, containing long chains of identical units.

The writer wishes to record his gratitude to Professor C. K. Ingold for his encouragement and advice, to the Dunlop Rubber Company for permission to publish this paper, and to Mr L. A. Wragg for assistance with the experimental part.

REFERENCES

- Farmer, E. H. 1946 In *Advances in colloid science*, p. 302. New York: Interscience.
- Fisher, H. L. & McColm, E. M. 1927 *Industr. Engng. Chem.* **19**, 1328.
- Flory, P. 1939 *J. Amer. Chem. Soc.* **61**, 1518.
- Gordon, M. 1948 *J. Pol. Sci.* **3**, 438.
- Gordon, M. 1949 Hevea rubber—Cyclization kinetics and proton affinity. Thesis, London University.
- Gordon, M. 1950 'The kinetics of cyclization of Hevea latex.' International Meeting of the Division of Rubber Chemistry, Cleveland. *Industr. Engng. Chem.*, (in the Press).
- Le Bras, J. & Delalande, A. 1950 *Les Dérivés Chimiques du Caoutchouc Naturel*, p. 39. Paris: Dunod.
- Marvel, C. S. & Levesque, C. L. 1939 *J. Amer. Chem. Soc.* **61**, 1682, 3234.
- Wall, F. T. 1940 *J. Amer. Chem. Soc.* **62**, 803.

INDEX TO VOLUME 204 (A)

- Adams, C. A. Two new streak cameras, 19.
- Allan, D. L. & Poole, M. J. Experiments on the reaction $T(d, n)^4\text{He}$. I. The 90° cross-section, 488.
- Allan, D. L. & Poole, M. J. Experiments on the reaction $T(d, n)^4\text{He}$. II. The angular distribution, 500.
- Andrade, E. N. da C. & Dodd, C. The effect of an electric field on the viscosity of liquids. II, 449.
- Ashmore, P. G. & Norrish, R. G. W. A study of sensitized explosions X. The kinetics of decomposition of chloropierin and of the hydrogen-oxygen and hydrogen-chlorine reactions sensitized by chloropierin, 34.
- Aust, K. T. & Chalmers, B. Surface energy and structure of crystal boundaries in metals, 359.
- Bagguley, D. M. S. & Griffiths, J. H. E. Paramagnetic resonance in chromic sulphate alums at room temperature, 188.
- Bamford, C. H., Bamford, Daphne & Collins, J. R. Kinetic studies on carbohydrates in alkaline conditions. II. The kinetics of the rearrangements of glucose and fructose in alkaline solution, 85.
- Bamford, C. H. & Collins, J. R. Kinetic studies on carbohydrates in alkaline conditions. I. The kinetics of the autoxidation of glucose, 62.
- Bamford, Daphne, *See* Bamford, Bamford & Collins.
- Baskett, A. C. & Hinshelwood, Sir Cyril. Oxidation-reduction processes in cultures of bacteria. I. The reducing power of *Bact. lactis aerogenes* (Abstract), 295.
- Binding energy of triton (Hu & Hsu), 476.
- Bleaney, B. Paramagnetic resonance spectra of five chromic sulphate alums at low temperatures, 203.
- Bleaney, B. Thermal properties of potassium chromic alum between 0.05 and 1°K , 216.
- Bowden, F. P. The initiation of an explosion and its growth to detonation, 20.
- Bowers, R. & Mendelssohn, K. The viscosity of liquid helium between 2 and 5°K , 366.
- Calculation of deflexions of the vertical from gravity anomalies (Cook), 374.
- Carbohydrates in alkaline solutions (Bamford, Bamford & Collins), 85; (Bamford & Collins), 62.
- Cave, L., Corner, J. & Liston, R. H. A. The scattering of gamma-rays in extended media. I. Perpendicular incidence on a plane slab, 223.
- Chalmers, B. *See* Aust & Chalmers.
- Chandrasekhar, S. The invariant theory of isotropic turbulence in magneto-hydrodynamics, 435.
- Chromic sulphate alums (Bleaney), 203.
- Collins, J. R. *See* Bamford, Bamford & Collins; Bamford & Collins.
- Compressibility and absorption frequency of ionic crystals (Szigeti), 51.
- Cook, A. H. The calculation of deflexions of the vertical from gravity anomalies, 374.
- Corner, J. *See* Cave, Corner & Liston.
- Corner, J., Day, F. A. G. & Weir, R. E. The scattering of gamma-rays in extended media. III. Problems with spherical symmetry, 329.
- Corner, J. & Liston, R. H. A. The scattering of gamma-rays in extended media. II. Back-scattering of gamma-rays from a thick slab, 323.
- Cosmic rays, gravitational theory of origin (Gething), 278.
- Courtney-Pratt, J. S. A new photographic method for studying fast transient phenomena, 27.

- Crank, J. & Robinson, C. Interferometric studies in diffusion. II. Influence of concentration and orientation on diffusion in cellulose acetate, 549.
- Critical examination of the gravitational theory of the origin of cosmic rays (Gething), 278.
- Crystal boundaries in metals (Aust & Chalmers), 359.
- Davies, R. M., Owen, J. D., Edwards, D. H. & Thomas, D. E. Pressure measurements in detonating gases, 17.
- Day, F. A. G. *See* Corner, Day & Weir.
- Detonation (discussion) (Penney and others), 1.
- Diffraction image (Wolf), 533.
- Diffusion, interferometric studies (Robinson), 339; (Crank & Robinson), 549.
- Discussion on detonation (Penney and others), 1.
- Dodd, C. *See* Andrade & Dodd.
- Dyson, J. An interferometer microscope, 170.
- Edwards, D. H. *See* Davies, Owen, Edwards & Thomas.
- Effect of an electric field on the viscosity of liquids. II (Andrade & Dodd), 449.
- Egerton, Sir Alfred & Warren, D. R. Kinetics of the hydrogen/oxygen reaction. I. The explosion region in boric acid-coated vessels, 465.
- Elementary particles, quantum theory (Roberts), 123.
- Essen, L. The velocity of propagation of electromagnetic waves derived from the resonant frequencies of a cylindrical cavity resonator, 260.
- Evans, W. M. Some characteristics of detonation, 12.
- Experimental and thermodynamic investigation of the hydrogen-titanium system (McQuillan), 309.
- Experiments on the reaction $T(d, n)^4\text{He}$. I. The 90° cross-section (Allan & Poole), 488.
- Experiments on the reaction $T(d, n)^4\text{He}$. II. The angular distribution (Allan & Poole), 500.
- Fading of detonation in solid explosives (Gurton), 31.
- Fatty acid and soap films (Menter & Tabor), 514.
- Fermi-Dirac functions of integral order (Rhodes), 396.
- Flowers, B. H. The theory of the T + D reaction, 503.
- Gamma-rays, scattering (Cave, Corner & Liston), 223; (Corner & Liston), 323; (Corner, • Day & Weir), 329.
- Gething, P. J. D. A critical examination of the gravitational theory of the origin of cosmic rays, 278.
- Gordon, M. Statistical kinetics of Hevea rubber cyclization, 569.
- Griffiths, J. H. E. *See* Bagguley & Griffiths.
- Gurton, O. A. The fading of detonation in solid explosives, 31.
- Helium, viscosity of liquid (Bowers & Mendelssohn), 366.
- Hevea rubber cyclization (Gordon), 569.
- High and low detonation velocity regimes in condensed explosives (Taylor), 30.
- Hinshelwood, Sir Cyril. *See* Baskett & Hinshelwood.
- Hsu, K-N. *See* Hu & Hsu.
- Hu, T-M. & Hsu, K-N. Binding energy of the triton, 476.
- Huang, K. & Rhys, Avril. Theory of light absorption and non-radiative transitions in F-centres, 406.
- Hulm, J. K. The thermal conductivity of tin, mercury, indium and tantalum at liquid helium temperatures, 98.
- Hydrogen/oxygen reaction, kinetics (Egerton & Warren), 465.
- Hydrogen-titanium system (McQuillan), 309.

- Initiation of an explosion and its growth to detonation (Bowden), 20.
- Interferometer microscope (Dyson), 170.
- Interferometric studies in diffusion. I. Determination of concentration distributions (Robinson), 339.
- Interferometric studies in diffusion. II. Influence of concentration and orientation on diffusion in cellulose acetate (Crank & Robinson), 549.
- Invariant theory of isotropic turbulence in magneto-hydrodynamics (Chandrasekhar), 435.
- Ionic crystals, compressibility and absorption frequency (Szigeti), 51.
- Jones, H. Theoretical considerations relating to the detonation of solid explosives, 9.
- Kinetic studies on carbohydrates in alkaline conditions. I. The kinetics of the autoxidation of glucose (Bamford & Collins), 62.
- Kinetic studies on carbohydrates in alkaline conditions. II. The kinetics of the rearrangements of glucose and fructose in alkaline solution (Bamford, Bamford & Collins), 85.
- Kinetics of the hydrogen/oxygen reaction. I. The explosion region in boric acid-coated vessels (Egerton & Warren), 465.
- Lambert, J. D. & Rowlinson, J. S. Ultrasonic dispersion in organic vapours, 424.
- Levy, M. Wave equations in momentum space, 145.
- Light distribution near focus in an error-free diffraction image (Wolf), 533.
- Liston, R. H. A. *See* Cave, Corner & Liston; Corner & Liston.
- McQuillan, A. D. An experimental and thermodynamic investigation of the hydrogen-titanium system, 309.
- Mendelssohn, K. *See* Bowers & Mendelssohn.
- Menter, J. W. & Tabor, D. Orientation of fatty acid and soap films on metal surfaces, 514.
- New photographic method for studying fast transient phenomena (Courtney-Pratt), 27.
- Norrish, R. G. W. *See* Ashmore & Norrish.
- Orientation of fatty acid and soap films on metal surfaces (Menter & Tabor), 514.
- Owen, J. D. *See* Davies, Owen, Edwards & Thomas.
- Oxidation-reduction processes in cultures of bacteria. I. The reducing power of *Bact. lactis aerogenes* (Abstract) (Baskett & Hinshelwood), 295.
- Paramagnetic resonance in chromic sulphate alums at room temperature (Bagguley & Griffiths), 188.
- Paramagnetic resonance spectra of five chromic sulphate alums at low temperatures (Bleaney), 203.
- Paterson, S. The reduction in reaction energy in cases of incomplete confinement, 30.
- Penney, W. G. and others. A discussion on detonation, 1.
- Physical chemistry of detonative reactions (Ubbelohde), 25.
- Poole, M. J. *See* Allan & Poole.
- Potassium chromic alums (Bleaney), 216.
- Pressure measurements in detonating gases (Davies, Owen, Edwards & Thomas), 17.
- Quantum theory of the elementary particles. I. Introduction and classical field dynamics (Roberts), 123.

- Reduction in reaction energy in cases of incomplete confinement (Paterson), 30.
- Rhodes, P. Fermi-Dirac functions of integral order, 396.
- Rhys, Avril. *See* Huang & Rhys.
- Roberts, K. V. On the quantum theory of the elementary particles. I. Introduction and classical field dynamics, 123.
- Robinson, C. *See* Crank & Robinson.
- Robinson, C. Interferometric studies in diffusion. I. Determination of concentration distributions, 339.
- Rowlinson, J. S. *See* Lambert & Rowlinson.
- Sanders, J. V. & Tabor, D. Structure of thin films of aliphatic esters and alcohols on metals, 525.
- Scattering of gamma-rays in extended media. I. Perpendicular incidence on a plane slab (Cave, Corner & Liston), 223.
- Scattering of gamma-rays in extended media. II. Back-scattering of gamma-rays from a thick slab (Corner & Liston), 323.
- Scattering of gamma-rays in extended media. III. Problems with spherical symmetry (Corner, Day & Weir), 329.
- Sensitized explosions. X (Ashmore & Norrish), 34.
- Similarity solutions to problems involving gas flow and shock waves (Taylor), 8.
- Some characteristics of detonation (Evans), 12.
- Some comments on the contribution by Dr Courtney-Pratt (Ubbelohde), 29.
- Stability of a fluid heated from below (Sutton), 297.
- Statistical kinetics of Hevea rubber cyclization (Gordon), 569.
- Structure of thin films of aliphatic esters and alcohols on metals (Sanders & Tabor), 525.
- Study of sensitized explosions. X. The kinetics of decomposition of the hydrogen-oxygen and hydrogen-chlorine reactions sensitized by chloropicrin (Ashmore & Norrish), 34.
- Surface energy and structure of crystal boundaries in metals (Aust & Chalmers), 359.
- Sutton, O. G. On the stability of a fluid heated from below, 297.
- Szigeti, B. Compressibility and absorption frequency of ionic crystals, 51.
- Tabor, D. *See* Menter & Tabor; Sanders & Tabor.
- Taylor, Sir Geoffrey. Similarity solutions to problems involving gas flow and shock waves, 8.
- Taylor, J. High and low detonation velocity regimes in condensed explosives, 30.
- Theoretical considerations relating to the detonation of solid explosives (Jones), 9.
- Theory of light absorption and non-radiative transitions in *F*-centres (Huang & Rhys), 406.
- Theory of the T + D reaction (Flowers), 503.
- Thermal conductivity of tin, mercury, indium and tantalum at liquid helium temperatures (Hulm), 98.
- Thermal properties of potassium chromic alum between 0.05 and 1° K (Bleaney), 216.
- Thomas, D. E. *See* Davies, Owen, Edwards & Thomas.
- Triton, binding energy (Hu & Hsu), 476.
- Two new streak cameras (Adams), 19.
- Ubbelohde, A. R. The physical chemistry of detonative reactions, 25. Some comments on the contribution by Dr Courtney-Pratt, 29.
- Ultrasonic dispersion in organic vapours (Lambert & Rowlinson), 424.
- Velocity of propagation of electromagnetic waves derived from the resonant frequencies of a cylindrical cavity resonator (Essen), 260.

Viscosity of liquid helium between 2 and 5° K (Bowers & Mendelssohn), 366.

Viscosity of liquids affected by an electric field (Andrade & Dodd), 449.

Warren, D. R. *See* Egerton & Warren.

Wave equations in momentum space (Levy), 145.

Weir, R. E. *See* Corner, Day & Weir.

Wolf, E. Light distribution near focus in an error-free diffraction image, 533.

END OF THE TWO HUNDRED AND FOURTH VOLUME (SERIES A)

25 JAN. 1955

